THE AMERICAN MATHEMATICAL MONTHLY

Volume 104, Number 1 January 1997

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and notes should be sent to the Editor at the MONTHLY's Utah office:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies of the complete manuscript (including all figures with captions and lettering), typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper.

Letters to the Editor on any topic, for publication or for private reading, are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

See the MONTHLY section of MAA Online for current information such as contents of issues, descriptive summaries of forthcoming articles, tips for authors, and preparation of manuscripts in TEX:
http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all material, typewritten on only one side of the paper.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Mr. JOE WATSON, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Reading the Master: Newton and the Birth of Celestial Mechanics

Bruce Pourciau

Dedicated to J. Bruce Brackenridge

One factor that has remained constant through all the twists and turns of the history of physical science is the decisive importance of the mathematical imagination.
—Freeman J. Dyson

1. In January of 1684, the young astronomer Edmund Halley travelled from Islington up to London for a meeting of the Royal Society. Later, perhaps over tea and chocolate at a nearby coffee house, he chatted casually about natural philosophy and other topics with Sir Christopher Wren and Robert Hooke. Talk soon turned to celestial motions, and Halley later reconstructed the conversation [**22**, p. 26]:

> I, having from the consideration of the sesquialter proportion of Kepler concluded that the centripetall force [to the Sun] decreased in the proportion of the squares of the distances reciprocally, came one Wednesday to town, where I met with S^{r} Christ. Wren and M^{r} Hook, and falling in discourse about it, M^{r} Hook affirmed that upon that principle all the Laws of the celestiall motions were to be demonstrated, and that he himself had done it. I declared the ill success of my attempts; and S^{r} Christopher to encourage the Inquiry said that he would give M^{r} Hook or me 2 months time to bring him a convincing demonstration thereof, and besides the honour, he of us that did it, should have from him a present of a book of 40 shillings. M^{r} Hook then said that he would conceale [his] for some time that other triing and failing, might know how to value it, when he should make it publick. ... I remember S^{r} Christopher was little satisfied that he could do it, and though M^{r} Hook then promised to show it him, I do not yet find that in that particular he has been as good as his word.

The two month deadline passed. Wren and Halley waited through the summer, but still the promised proof from Hooke never came. Finally, in August, Halley would wait on Hooke no longer. He carried the question to Cambridge and the Lucasian Professor of Mathematics, Isaac Newton.

Newton's secretary and attendant has painted a portrait, daubed with colorful and concrete detail, of the eccentric Cambridge professor Halley had finally decided to approach [**12**, p. xiii–xiv]:

> I cannot say, I ever saw him laugh, but once... I never knew him take any Recreation or Pastime, either in Riding out to take y^{e} Air, Walking, Bowling

or any other Exercise whatever, thinking all Hours lost, y^t was not spent in his Studyes, to w^{ch} he kept so close ... so intent, so serious upon [them], y^t he eat very sparingly, nay, oft times he has forgot to eat at all, so y^t going into his Chamber I have found his Mess untouch'd, of w^{ch} when I have reminded him, [he] would reply, Have I; & then making to y^e Table, would eat a bit or two standing, for I cannot say, I ever saw Him sit at Table by himself ... He very rarely went to Dine in y^e Hall unless upon some Publick Dayes, & then, if He has not been minded, would go very carelessly, w^{th} Shooes down at Heels, Stockins unty'd, Suplice on, & his Head scarcely comb'd.... At some seldom Times when he design'd to dine in y^e Hall [he] would turn to y^e left hand, & go out into y^e street, where making a Stop, when he found his mistake, [he] would hastily turn back & then sometimes instead of going into y^e Hall, would return to his Chamber again....

... in his Garden, w^{ch} was never out of Order, ... he would, at some seldom Times, take a short Walk or two, not enduring to see a Weed in it.... When he has some Times taken a turn or two [he] has made a sudden Stand, turn'd himself about, run up y^e Stairs [&] like another A[r]chimides, with an *ευρηκα* fall to write on his Desk standing, without giving himself the Leasure to draw a Chair to sit down on....

In a letter from 1727 [**22**, p. 27], Abraham de Moivre set the scene as Halley, having arrived in Cambridge, posed the crucial question to the reclusive mathematician:

... after they had been some time together, the D^r asked [Newton] what he thought the Curve would be that would be described by the Planets supposing the force of attraction towards the Sun to be reciprocal to the square of their distance from it. S^r Isaac replied immediately that it would be an Ellipsis. The Doctor struck with joy and amazement asked him how he knew it. Why saith he I have calculated it....

Witness the birth of celestial mechanics: the embryonic question has been answered—

every orbital motion subject to an inverse-square force lies on a conic having focus at the force center

—not with a guess, but with a *mathematical demonstration*!

Semester after semester, at every college and university, we give our students the same answer Newton gave to Halley, our demonstrations—so different from Newton's—blessed by the glories of vector calculus, and in this way we honor Newton and celebrate the emergence of celestial dynamics. In the present article, we honor Newton in the way of Abel, who counsels us to read the masters. We shall place the original argument from Newton's *Principia* next to a modern counterpart, delighting in the stark contrasts. One delightful difference: Newton's argument requires that we first answer the *converse* to Halley's question—

What force law maintains a conic motion orbiting about the focus?

—and again, reading the master, we shall juxtapose the *Principia*'s very geometric proof of this reversal with its demonstration by vector calculus. In this mix of old

and new, of geometry and analysis, some insights and surprises make their way to the surface:

- The mathematics of the *Principia* is geometric *analysis*, both analysis in the sense of 'taking apart' as well as analysis in the sense of *calculus*. Newton's geometry is calculus—limits, derivatives, integrals, acceleration, curvature—masked as geometry.
- While less precise than their vector calculus descendants, the *Principia*'s definitions have a concrete, visceral character that informs our geometric and physical intuition.
- The first ten sections of the *Principia* (apart from the statement of the Third Law) contain no physics, only mathematics. Newton may write of 'forces,' but he calculates accelerations. His concentration on acceleration and shape reminds us that force and mass take no part in the mathematics of the one-body problem, which occupies the leading sections of the *Principia*.
- In contrast to force, curvature is deeply involved with the *Principia*'s orbital dynamics, yet apart from rare oblique sightings, the dependence on curvature remains hidden.
- Asked who should receive credit for answering Halley's question with a demonstration rather than a guess, historians of science bow to Newton. Asked for evidence to back up their claim, the historians open the *Principia* and point to a *two-sentence* argument. We confirm that Newton's little sketch, given air and sun, blossoms into a cogent proof.
- Reading the masters—Archimedes, Newton, Euler, Gauss, Riemann,... —can mean entering a foreign paradigm, an unfamiliar mathematical world where alien values, language, definitions, tools, strategies, and assumptions frustrate our attempts to understand. And so it is with the *Principia*. But with persistence and prayer, even the *Principia* sends up her secrets. As we slowly learn to navigate in Newton's world, we deepen our understanding of the *Principia*'s paradigm as well as our own.

It may seem odd to have placed our conclusions here in the introduction, but with these closing remarks now out of the way, we can read on unburdened by the western need to fret and fuss about the point of it all. As the Taoist philosopher Chuang Tzu suggests [**19**, p. 126], we can now lean back and float with the current, "going under with the swirls and coming out with the eddies, following along the way the water goes, and never thinking... ."

2. We begin with Newton's generalized answer to Halley—that every orbit produced by an inverse-square force must lie on a conic—in this section giving a contemporary proof and in the next exploring the *Principia*'s original argument. But we should first agree on some technical vocabulary, so that we can be more precise. Any smooth map $\mathbf{r} = \mathbf{r}(t)$ from an open interval J into euclidean 3-space is a *motion*. Every motion $\mathbf{r}$ has a velocity $\mathbf{v} = \dot{\mathbf{r}}$ and an acceleration $\mathbf{a} = \dot{\mathbf{v}}$. For the magnitude of a vector, we choose the same letter in nonbold italic: thus, for example, $r = |\mathbf{r}|$, $v = |\mathbf{v}|$, and $a = |\mathbf{a}|$. (We tacitly assume that r and v (the *speed*) never vanish.) We say the motion $\mathbf{r}$ has an *inverse-square acceleration* provided for some nonzero λ,

$$\mathbf{a} = \frac{-\lambda}{r^2}\mathbf{U}$$

for all t in J. Here $\mathbf{U}$ stands for the unit direction vector $\mathbf{r}/r$. More generally, whenever the cross-product $\mathbf{r} \times \mathbf{a}$ vanishes identically, we call $\mathbf{r}$ an *orbital motion*.

If the origin S has some significance—it might be the focus of a conic or the pole of a spiral, for instance—an orbital motion may be labelled a *motion about* S. A sentence that would be typical of the *Principia*, "A body is urged by a centripetal force continually directed toward an immovable center S," becomes briefer in our language: "Given a motion about S."

Assuming that Mars traversed an ellipse with its position vector sweeping out equal areas in equal times, Kepler made predictions in his *Astronomica nova* of 1609 that matched the careful observations of Tycho Brahe. In Propositions I and II (Section II, Book I) of the *Principia*, Newton uses this area principle to characterize orbital motions in general [**11**, p. 40 and 42]:

PROPOSITION I THEOREM I

The areas which revolving bodies describe by radii drawn to an immovable centre of force do lie in the same immovable planes, and are proportional to the times in which they are described.

PROPOSITION II THEOREM II

Every body that moves in any curved line described in a plane, and by a radius drawn to a point either immovable, or moving forwards with an uniform rectilinear motion, describes about that point areas proportional to the times, is urged by a centripetal force directed to that point.

Today of course we translate these propositions into the language of vectors:

NEWTON'S AREA THEOREM *For any motion* $\mathbf{r} = \mathbf{r}(t)$, *the following are equivalent:*

(a) $\mathbf{r}$ *is orbital*

(b) *the (massless) angular momentum* $\mathbf{h} = \mathbf{r} \times \mathbf{v}$ *is constant*

(c) $\mathbf{r}$ *is planar and sweeps out area at a constant rate*

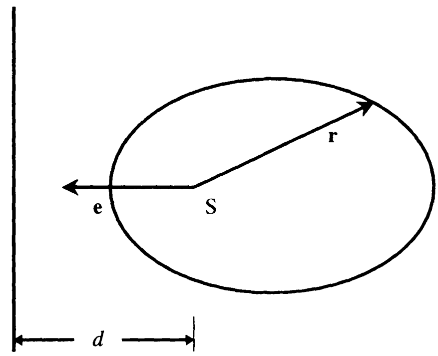
The proof is simple, especially once we agree that the area swept out is

$$\frac{1}{2}\int_{t_0}^{t} |\mathbf{r} \times \mathbf{v}|\, dt,$$

the only slippery step being to show $\mathbf{r}$ is planar when $\mathbf{h}$ vanishes everywhere, but in this case the derivative $\dot{\mathbf{U}}$ vanishes everywhere (recall $\mathbf{U} = \mathbf{r}/r$), indicating that the motion lies on a fixed ray from the origin. That $\dot{\mathbf{U}}$ remains zero follows from a simple fact:

$$\dot{\mathbf{U}} = \frac{\mathbf{h} \times \mathbf{r}}{r^3} \tag{1}$$

Halley's question and Newton's answer involve the relationship between the acceleration of the motion and the shape of the orbit. Moving from acceleration to shape, we define the *trajectory* of a motion $\mathbf{r} = \mathbf{r}(t)$ to mean the subset $\{\mathbf{r}(t)\colon t \in J\}$ of 3-space. An *orbit* is then just the trajectory of an orbital motion. If a trajectory lies on a conic, say, or a spiral, we would have a *conic* or *spiral motion*. The Principian sentence, "A body, urged by a centripetal force continually directed toward an immovable center S, moves in a conic section with focus at S," now turns into "Consider a conic motion about S." Of course conics hold some special interest for us here, and we recall the following definition: a *conic* is the locus of points whose distance from a given point S (the *focus*) is some positive constant e



(the *eccentricity*) times the distance from a given line (the *directrix*). Perhaps we should put this definition in vector dress, so it will feel more comfortable when vector calculus comes to call. If we let $\mathbf{r}$ be the position vector from the focus, d the distance from the directrix to the focus, and $\mathbf{e}$ (the *eccentricity vector*) a vector of length e which points perpendicularly toward the directrix, then the definition tells us that

$$r = e\left(d - \mathbf{r} \cdot \frac{\mathbf{e}}{e}\right),$$

and with the notation $\mathbf{U} = \mathbf{r}/r$ and $l = de$, this formula turns into the *vector conic equation*:

$$\mathbf{r} \cdot (\mathbf{e} + \mathbf{U}) = l. \tag{2}$$

The constant l is called the *semi-latus rectum* of the conic. Given a positive constant l and a nonzero vector $\mathbf{e}$, the vector conic equation defines a conic with semi-latus rectum l, eccentricity $e = |\mathbf{e}|$, axis along $\mathbf{e}$, and focus at the origin. When $\mathbf{e} = \mathbf{0}$, then (2) describes a circle of radius l about the origin, and if $l = 0$, we have a ray from the origin.

At this point, we have the vocabulary and background to explore a contemporary version of Newton's answer to Halley. Suppose we have a motion $\mathbf{r} = \mathbf{r}(t)$ with an inverse-square acceleration, so that for some nonzero number λ,

$$\mathbf{a}(t) = \frac{-\lambda}{r^2}\mathbf{U}(t)$$

for all t in some open interval J. Crossing with the angular momentum $\mathbf{h} = \mathbf{r} \times \mathbf{v}$, we have

$$\begin{aligned} \mathbf{a} \times \mathbf{h} &= \frac{-\lambda}{r^2}\mathbf{U} \times \mathbf{h} \\ &= -\lambda \frac{\mathbf{r} \times \mathbf{h}}{r^3} \end{aligned}$$

which becomes, using (1),

$$\mathbf{a} \times \mathbf{h} = \lambda \dot{\mathbf{U}}.$$

Now antidifferentiate, remembering that $\mathbf{h}$ is constant because $\mathbf{r}$ is orbital:

$$\begin{aligned} \mathbf{v} \times \mathbf{h} &= \lambda \mathbf{U} + \mathbf{c} \\ &= \lambda(\mathbf{U} + \mathbf{e}) \end{aligned}$$

for some constant vectors $\mathbf{c}$ and $\mathbf{e} = \frac{1}{\lambda}\mathbf{c}$. If we dot with $\mathbf{r}$, we find

$$\frac{1}{\lambda}\mathbf{r}\cdot(\mathbf{v}\times\mathbf{h}) = \mathbf{r}\cdot(\mathbf{e}+\mathbf{U}),$$

and then permuting the entries in the scalar triple product uncovers the vector conic equation (2):

$$\frac{h^2}{\lambda} = \mathbf{r}\cdot(\mathbf{e}+\mathbf{U}).$$

When the constant vector $\mathbf{h}$ vanishes, this reduces to $\mathbf{U} = -\mathbf{e}$, and the motion must then lie on a fixed ray from the origin. If $\mathbf{h}$ does not vanish, but $\mathbf{e}$ does, we conclude $r = h^2/\lambda$, so the orbit lies on a circle centered at the origin. Supposing neither $\mathbf{h}$ nor $\mathbf{e}$ vanishes, we have seen that the vector conic equation (2) defines a conic with focus at the origin. And that seals it:

NEWTON'S SHAPE THEOREM. *Apart from motion on a ray from the center, every motion with an inverse-square acceleration must be a conic motion about the focus.*

A second proof of the Shape Theorem is quick but sly. Assume again that

$$\mathbf{a}(t) = \frac{-\lambda}{r^2}\mathbf{U}(t)$$

Then of course $\mathbf{h}$ remains constant, but (surprise!) so does the vector $\mathbf{L} = \frac{1}{\lambda}\mathbf{v}\times\mathbf{h} - \mathbf{U}$. To check, compute the derivative:

$$\dot{\mathbf{L}} = \frac{1}{\lambda}\mathbf{a}\times\mathbf{h} - \frac{\mathbf{h}\times\mathbf{r}}{r^3} = \frac{1}{\lambda}\left(\frac{-\lambda}{r^2}\mathbf{U}\right)\times\mathbf{h} - \frac{\mathbf{h}\times\mathbf{U}}{r^2} = \mathbf{0}$$

Now just dot $\mathbf{r}$ with $\mathbf{L}+\mathbf{U}$,

$$\mathbf{r}\cdot(\mathbf{L}+\mathbf{U}) = \frac{1}{\lambda}\mathbf{r}\cdot(\mathbf{v}\times\mathbf{h}) = \frac{h^2}{\lambda},$$

and we recognize the vector conic equation (2). That's all there is to it.

The sly part of this proof is (un)clear: why would one *expect* the vector $\frac{1}{\lambda}\mathbf{v}\times\mathbf{h} - \mathbf{U}$ to be constant? The secret lies in a formula for the eccentricity vector $\mathbf{e}$. Given any conic motion $\mathbf{r} = \mathbf{r}(t)$, if we differentiate the vector conic equation,

$$\mathbf{r}\cdot(\mathbf{e}+\mathbf{U}) = l,$$

and solve for the (constant) eccentricity vector $\mathbf{e}$, we obtain the

ECCENTRICITY FORMULA. *For any motion* $\mathbf{r} = \mathbf{r}(t)$ *satisfying the vector conic equation* (2),

$$\mathbf{e} = \frac{l}{h^2}\mathbf{v}\times\mathbf{h} - \mathbf{U}. \tag{3}$$

Of course we began with an inverse-square motion, not a conic motion, but if we had had a conic motion, then the vector $(l/h^2)\mathbf{v}\times\mathbf{h} - \mathbf{U}$, representing as it does the eccentricity vector, would have been *a priori* constant. Knowing that λ turns out to be h^2/l (see our first proof), it seems natural then to suspect that $\mathbf{L} = (1/\lambda)\mathbf{v}\times\mathbf{h} - \mathbf{U}$ should be constant in the case of inverse-square acceleration. If you do not like this sneaky proof of the Shape Theorem, blame Laplace. The vector $\mathbf{L}$, sometimes called the Laplace-Runge-Lenz vector, has the history of its rediscoveries etched in its name.

Now that we have seen two contemporary proofs, let us drift back in time, back to the 1680s, to examine Newton's original argument for the Shape Theorem in the *Principia*.

3. Only with some nervousness, do we open Newton's monumental work *Philosophiae Naturalis Principia Mathematica*. It had a reputation in 1687; it has a reputation still—a reputation for being impenetrable. In the latter half of the eighteenth century and on into the nineteenth, this reputation fed a cottage industry of writing notes and commentaries devoted entirely to 'understanding' the *Principia*. (The industry may have declined, but it still produces excellent commentaries from time to time: witness [**5**] and [**6**], just out in 1995.) Always formal, terse, and crabbed in his scholarly work, Newton took these stylistic tendencies to their limit in the *Principia*. Why? A decade earlier, his theory of colors had been attacked by Leibniz, Hooke, Linus, Lucas, as well as others, and Newton had detested the controversy. In a shrill letter to Henry Oldenburg, who was then Secretary of the Royal Society, Newton despairs, "I see I have made myself a slave to Philosophy, but if I get free of Mr. Linus's business I will resolutely bid adew to it eternally, excepting what I do for my private satisfaction or leave to come out after me. For I see a man must either resolve to put out nothing new or become a slave to defend it." [**7**, p. 198] Of course, Newton did not "leave [the *Principia*] to come out after [him]," but he did choose to limit his readership and therefore his potential critics by composing in an icy, mathematical style, ultimately producing 500 pages of dense Latin text—definitions, axioms, lemmas, theorems, propositions, demonstrations, scholia, and figures, all fixed in place, a massive ordered regiment of abstract formality. According to a close friend of Newton's [**2**, p. 168], controversy of any kind

> made sr Is[aac] very uneasy; who abhorred all Contests... And for this reason, mainly to avoid being baited by little Smatterers in Mathematicks, he told me, he designedly made his Principia abstruse; but yet so as to be understood by able Mathematicians, who he imagined, by comprehending his Demonstrations, would concurr with him in his Theory.

Yet even the most able mathematicians of the day struggled with the *Principia*. The confident young mathematician Abraham de Moivre happened to be visiting the Duke of Devonshire when Newton arrived to present the Duke with a copy of the new work [**21**, p. 471]:

> [de Moivre] opened the book and deceived by its apparent simplicity persuaded himself that he was going to understand it without difficulty. But he was surprised to find it beyond the range of his knowledge and to see himself obliged to admit that what he had taken for mathematics was merely the beginning of a long and difficult course that he had yet to undertake. He purchased the book, however; and since the lessons he had to give forced him to travel about continually, he tore out the pages in order to carry them in his pocket and to study them during his free time.

Prepared by its scary reputation, we cannot conjure up the initial poise of de Moivre as we open the *Principia*, but prepared for some hard work, let us take a look at Newton's argument for the Shape Theorem. Actually, to do this in the proper order, we should close the *Principia* for the moment and begin nearer the

beginning, returning to Halley's call on Newton in 1684. Earlier we have read de Moivre's description of their meeting [**22**, p. 27]:

> ... after they had been some time together, the D^r asked him what he thought the Curve would be that would be described by the Planets supposing the force of attraction towards the Sun to be reciprocal to the square of their distance from it. S^r Isaac replied immediately that it would be an Ellipsis. The Doctor struck with joy and amazement asked him how he knew it. Why saith he I have calculated it... .

But stopping here is a rude interruption, for de Moivre continues [**7**, p. 283],

> ... whereupon D^r Halley asked him for his calculation without any farther delay, S^r Isaac looked among his papers but could not find it, but he promised him to renew it, & sent it.

It would be three months before Newton made good his promise, but idleness had not caused the delay, for he not only renewed his calculation for the ellipse, but embedded that calculation in a nine-page tract, "De motu Corporum in gyrum" ("On the Motion of Bodies in Orbit"), which Halley received in November.

It is in "De motu" then that we should look for Newton's original demonstration of the Shape Theorem, that an inverse-square force implies conic orbits. Thumbing through its pages, we pass a line of definitions, hypotheses, theorems, corollaries, and problems until we stop at a familiar-looking claim [**12**, VI p. 49]:

> Scholium *The major planets orbit, therefore, in ellipses having a focus at the centre of the Sun ... exactly as Kepler supposed.*

The Shape Theorem (at least for ellipses)! Eagerly we anticipate the proof—hunched over the scholium, eyes narrowed, pencil poised—but then the adrenaline seeps away as we scan down the page to find ... nothing. Newton has left the Shape Theorem, his answer to Halley, as a bald claim, completely unsupported! Because the scholium directly follows

> Problem 3 *A body orbits in an ellipse: there is required the law of centripetal force tending to a focus of the ellipse.*

we would guess that Newton must have viewed the Shape Theorem as a trivial corollary of his solution to Problem 3, or, more generally, of what we shall call

NEWTON'S ACCELERATION THEOREM. *Every conic motion about the focus has an inverse-square acceleration.*

Not understanding how the Shape Theorem would follow trivially from the Acceleration Theorem, we turn from "De motu" to the *Principia*, expecting the fuller development there to enlighten us.

Halley's question in August of 1684 had reseeded Newton's interest in celestial mechanics, and "De motu" was just the first little sprout. In January of 1685, he wrote Flamsteed, the Astronomer Royal, "Now that I am upon this subject, I would gladly know ye bottom of it before I publish my papers." [**7**, p. 286] What understatement: between November of 1684 and April of 1687, Newton came to

"know ye bottom of it," and the nine-page treatise exploded into a five hundred page masterpiece.

Now remember that "De motu" had left the Shape Theorem unproved. And the 1687 *Principia*? No better! In Section III of Book I, Newton demonstrates Propositions XI–XIII, which, taken together, establish the Acceleration Theorem and then follows with the Shape Theorem dressed as a corollary [**11**, p. 61] to this trio of propositions:

> Cor. I *From the three last Propositions it follows, that if a body* P *goes from place* P *with any velocity in the direction of any right line* PR, *and at the same time is urged by the action of a centripetal force that is inversely proportional to the square of the distance of the places from the center, the body will move in one of the conic sections, having its focus in the center of force*... .

But again, no proof. Worse yet, no one complained—not Halley, not Leibniz, not Huygens, not de Moivre—until, in October of 1710, *twenty-three years* after the publication of the *Principia*, Johann Bernoulli finally pointed out the obvious: Corollary I needed a demonstration. By this time, however, perhaps getting an early wind of Bernoulli's criticism, Newton had already decided to fill the gap, instructing his editor, in a letter dated 11 October 1709, to slip the following argument [**13**, p. 5–6] into the second edition (1713) of the *Principia*:

> *Nam datis umbilico et puncto contactus & positione tangentis, describi potest Sectio conica quae curvaturam datam ad punctum illud habebit. Datur autem curvatura ex data vi centripeta: et Orbes duo se mutuo tangentes eadem vi describi non possunt.*

For the third edition (1726), Newton added to this shockingly brief sketch the word 'velocity' in two places, resulting in [**11**, p. 61]

NEWTON'S ARGUMENT FOR THE SHAPE THEOREM

> *For the focus, the point of contact, and the position of the tangent, being given, a conic section may be described, which at that point shall have a given curvature. But the curvature is given from the centripetal force and velocity of the body being given; and two orbits, touching one the other, cannot be described by the same centripetal force and the same velocity.*

Brevity may be the soul of wit, but it may be the seed of confusion as well. No one laughs when a fundamental proposition of celestial mechanics is followed by a two-sentence sketch which fails to persuade. At least Newton's *plan*, although strikingly different from what we saw in Section 2, seems both familiar and clear—to prove that every solution to a given initial-value problem has a particular form, we exhibit a solution of that form and then invoke a uniqueness principle—but connecting all the dots in the outline may be another story, especially when some of the dots themselves are missing.

Expanding Newton's sketch in a natural way, we arrive at what we take as his intended strategy:

NEWTON'S STRATEGY FOR PROVING THE SHAPE THEOREM

1. Suppose given any motion $\bar{\mathbf{r}} = \bar{\mathbf{r}}(t)$ with an inverse-square acceleration. At some time t_0, note the position $\mathbf{r}_0$, velocity $\mathbf{v}_0$, and curvature κ_0 of the motion $\bar{\mathbf{r}}$.

2. Construct a conic $\mathscr{C}$, having focus at the origin, that passes through the tip of $\mathbf{r}_0$ with tangent parallel to $\mathbf{v}_0$ and curvature κ_0.
3. On the conic $\mathscr{C}$, put a motion $\mathbf{r} = \mathbf{r}(t)$ about the focus that leaves the tip of $\mathbf{r}_0$ with velocity $\mathbf{v}_0$. (Newton never mentions this step, which involves making sure the position vector sweeps out area at a uniform rate, but it's a simple matter, and one that he probably took for granted.)
4. From Propositions XI–XIII (the Acceleration Theorem), infer that $\mathbf{r} = \mathbf{r}(t)$, a conic motion about the focus, must have an inverse-square acceleration.
5. Thus both $\mathbf{r}$ and $\bar{\mathbf{r}}$ have inverse-square accelerations, but even better, the matching of position, velocity, and curvature is steps (2) and (3) forces $\mathbf{r}$ and $\bar{\mathbf{r}}$ to share the same proportionality constant.
6. Finally, noting that $\mathbf{r}$ and $\bar{\mathbf{r}}$ now both solve the same initial-value problem, invoke a uniqueness principle to conclude that $\mathbf{r} = \bar{\mathbf{r}}$, proving that our given inverse-square motion $\bar{\mathbf{r}}$ must be a conic motion about the focus as desired.

As we begin to check whether this six-step strategy unfolds further into a convincing proof, we can see already that step (2) will block us, unless we know a little about the curvature of conics. For a motion $\mathbf{r} = \mathbf{r}(t)$, the curvature κ is $|\dot{\mathbf{T}}|/v$ and the radius of curvature ρ is $1/\kappa$, where $\mathbf{T}$ is the unit tangent $\mathbf{v}/v$. From the velocity and the acceleration, we can easily find the curvature from a well-known formula:

$$\rho = \frac{v^3}{|\mathbf{a} \times \mathbf{v}|}. \tag{4}$$

To calculate the radius of curvature for a conic, we start with any motion $\mathbf{r} = \mathbf{r}(t)$ satisfying the vector conic equation (2),

$$\mathbf{r} \cdot (\mathbf{e} + \mathbf{U}) = l,$$

differentiate twice to get

$$\mathbf{a} \cdot (\mathbf{e} + \mathbf{U}) + \mathbf{v} \cdot \frac{\mathbf{h} \times \mathbf{r}}{r^3} = 0,$$

and insert our formula (3) for the eccentricity vector $\mathbf{e}$ to see that

$$\frac{l}{h^2}\mathbf{a} \cdot (\mathbf{v} \times \mathbf{h}) + \mathbf{v} \cdot \frac{\mathbf{h} \times \mathbf{r}}{r^3} = 0.$$

Sliding the entries in the scalar triple products gives back

$$|\mathbf{a} \times \mathbf{v}| = \frac{1}{l}\left(\frac{h}{r}\right)^3,$$

which leads to

$$\rho = \frac{v^3}{|\mathbf{a} \times \mathbf{v}|} = l\left(\frac{rv}{h}\right)^3,$$

or, rephrasing, to the

CONIC CURVATURE LEMMA. *For any conic motion with semi-latus rectum l,*

$$\rho = \frac{l}{|\mathbf{U} \times \mathbf{T}|^3}. \tag{5}$$

Newton cast this lemma more elegantly [**12**, III p. 159]: *If the line perpendicular to the conic at* P *meets the focal axis at* N, *then ρ varies as* PN^3. (The equivalence to

our lemma follows from a geometric fact about conics: the projection of PN onto SP is the semi-latus rectum.) This lovely property is just one of several striking results on curvature obtained by Newton in his 1671 tract on series and fluxions. "The problem [of curvature]," he wrote in this tract, "has the mark of exceptional elegance and of being pre-eminently useful in the science of curves." [**12**, III p. 151] From an insight in his *Waste Book* made around December of 1664 (over twenty years before the *Principia*), we have evidence that Newton also recognized the fundamental place of curvature in the study of orbital motions: "If the body b moved in an Ellipsis, then its force in each point (if its motion in that point bee given) may be found by a tangent circle of equall crookedness [read curvature] with that point of the Ellipsis." [**22**, p. 14] It is perhaps surprising then that curvature plays no role in the 1687 *Principia*. However, in the 1690s Newton made radical plans for revising the first edition, plans that would have made curvature the centerpiece of his celestial mechanics. Sadly, this radical revision never made it into print, and in the end Newton contented himself with relatively minor changes, squeezing some curvature methods into the second (1713) and third (1726) editions as tacked on corollaries. For more on the role of curvature in Newton's celestial mechanics, see [**3**, **4**, **10**, and **17**].

Now that we know something about the curvature of conics, we can begin to connect all the dots in a proof of the Shape Theorem inspired by Newton's two-sentence argument in the *Principia*. We follow the six-step strategy above, for it seems to be the only plausible interpretation of what Newton had in mind.

Step 1: We give ourselves any motion $\bar{\mathbf{r}} = \bar{\mathbf{r}}(t)$ with an inverse-square acceleration: for some nonzero λ, suppose $\bar{\mathbf{r}}$ solves the initial-value problem

$$\ddot{\mathbf{r}}(t) = \frac{\lambda}{r^2}\mathbf{U}(t), \quad \mathbf{r}(t_0) = \mathbf{r}_0, \quad \dot{\mathbf{r}}(t_0) = \mathbf{v}_0$$

on the open interval J. If $\mathbf{r}_0 \times \mathbf{v}_0 = \mathbf{0}$, then the motion lies on a fixed ray through the origin, but apart from this special case, we need to prove that $\bar{\mathbf{r}}$ is a conic motion about the focus. Since $\bar{\mathbf{r}}$ is an orbital motion, the orbit lies in a fixed plane and the angular momentum remains fixed at $\mathbf{h}_0 = \mathbf{r}_0 \times \mathbf{v}_0$.

Step 2: In this fixed plane, we now construct a conic that "fits" the orbit of $\bar{\mathbf{r}}$. Let ρ_0 be the radius of curvature of $\bar{\mathbf{r}}$ at $\bar{\mathbf{r}}(t_0) = \mathbf{r}_0$. Put

$$l = \rho_0 |\mathbf{U}_0 \times \mathbf{T}_0|^3$$

$$\mathbf{e} = \frac{l}{h_0^2}\mathbf{v}_0 \times \mathbf{h}_0 - \mathbf{U}_0$$

where $\mathbf{U}_0 = \mathbf{r}_0/r_0$, $\mathbf{T}_0 = \mathbf{v}_0/v_0$, and $\mathbf{h}_0 = \mathbf{r}_0 \times \mathbf{v}_0$. (As $\mathbf{r}_0$ and $\mathbf{v}_0$ are not parallel, $\mathbf{h}_0 \neq 0$ and $\mathbf{e}$ is well-defined.) The vector-conic equation (2)

$$\mathbf{r} \cdot (\mathbf{e} + \mathbf{U}) = l$$

now defines a particular conic $\mathscr{C}$. One easily checks that $\mathscr{C}$ has a focus at the origin, and that $\mathscr{C}$ passes through the tip of $\mathbf{r}_0$ with its tangent parallel to $\mathbf{v}_0$ and its radius of curvature equal to ρ_0.

Step 3: At this point, we would like to apply Newton's Acceleration Theorem to our constructed conic, but the Acceleration Theorem applies only to conic *motions*, indeed only to conic motions about the focus, not to mere conic loci. Therefore, on

the conic locus $\mathscr{C}$ we now place a motion about the focus. (To put it differently, we must parameterize the conic locus $\mathscr{C}$ in a way that keeps the acceleration vector pointed at the focus.) By the Area Theorem, to make a motion about the focus, we need only make a motion whose position vector from the focus sweeps out area at a constant rate, and intuitively we can do this by arranging for the area swept out to be our parameter. More precisely: Using arc-length measured from the tip of $\mathbf{r}_0$, let $\mathbf{r}_1 = \mathbf{r}_1(s)$ be the unit-speed motion on $\mathscr{C}$ having initial velocity $\mathbf{T}_0$. The real function

$$a(s) = t_0 + \int_0^s \frac{1}{h_0} |\mathbf{r}_1(s) \times \dot{\mathbf{r}}_1(s)|\, ds$$

is smooth and strictly increasing. (Note that $h_0 = |\mathbf{r}_0 \times \mathbf{v}_0| \neq 0$ and $|\mathbf{r}_1(s) \times \dot{\mathbf{r}}_1(s)| \neq 0$ for all s, because tangents to $\mathscr{C}$ never pass through the focus.) Take the (smooth) inverse $a^{-1} = a^{-1}(t)$, and use it to define a motion $\mathbf{r} = \mathbf{r}(t)$ on $\mathscr{C}$ by

$$\mathbf{r}(t) = \mathbf{r}_1[a^{-1}(t)].$$

This constructed conic motion $\mathbf{r}$ is also a motion about the focus S, for it has constant angular momentum $\mathbf{h}_0 = \mathbf{r}_0 \times \mathbf{v}_0$. Moreover, $\mathbf{r}(t_0) = \mathbf{r}_0$ and $\dot{\mathbf{r}}(t_0) = \mathbf{v}_0$.

We haven't done anything here, by the way, that Newton couldn't do. You can find him geometrically constructing motions about the focus, on given conic loci, in the *Principia*, Book I, Section VI [**11**, p. 109–116]. Such constructions are even implicit in Newton's proof of the Area Theorem in Propositions I and II, at the very beginning of the *Principia*. In his two-sentence argument for the Shape Theorem, Newton fails to mention the problem of putting an orbital motion on his constructed conic, but at the *Principia*'s level of rigor, this is a trivial omission. Refer to [**15** and **16**] for some discussion of this point.

Step 4: We apply the Acceleration Theorem (Propositions XI–XIII, Section III, Book I) to $\mathbf{r} = \mathbf{r}(t)$, our newly minted conic motion about the focus, and conclude that $\mathbf{r}$ has an inverse-square acceleration: for some nonzero μ,

$$\ddot{\mathbf{r}}(t) = \frac{\mu}{r^2}\mathbf{U}(t)$$

for all t.

Step 5: To prove that $\mu = \lambda$, we return to the curvature matching we did in Step 2. By design, both our constructed motion $\mathbf{r}$ and our given motion $\bar{\mathbf{r}}$ share the same radius of curvature at the tip of $\mathbf{r}_0$, namely ρ_0. For the conic motion $\mathbf{r}$, by (4),

$$\rho_0 = \frac{v_0^3}{|\mathbf{a}_0 \times \mathbf{v}_0|} = \frac{v_0^3}{\left|\dfrac{\mu}{r_0^2}\mathbf{U}_0 \times \mathbf{v}_0\right|} = \frac{h_0^2}{\mu|\mathbf{U}_0 \times \mathbf{T}_0|^3}$$

Similarly, for the given motion $\bar{\mathbf{r}}$,

$$\rho_0 = \frac{h_0^2}{\lambda|\mathbf{U}_0 \times \mathbf{T}_0|^3}.$$

It follows that $\mu = \lambda$.

Step 6: We now have *two* solutions, the constructed conic motion $\mathbf{r}$ and the given inverse-square motion $\bar{\mathbf{r}}$, to the initial-value problem

$$\ddot{\mathbf{r}}(t) = \frac{\lambda}{r^2}\mathbf{U}(t), \quad \mathbf{r}(t_0) = \mathbf{r}_0, \quad \dot{\mathbf{r}}(t_0) = \mathbf{v}_0$$

on the interval J. By standard uniqueness theorems (equivalent to Propositions XLI and XLII, Section VIII, Book I, *Principia*) for differential equations, we conclude that $\mathbf{r} = \bar{\mathbf{r}}$ on J, and it follows that our given inverse-square motion must be a conic motion about the focus, as expected.

This completes a "Newtonian" proof of the Shape Theorem—that *every motion having an inverse-square acceleration is a conic motion about the focus*—a proof springing from Newton's two-sentence argument in the *Principia*. Is this proof the contemporary version of what Newton had in mind? Probably, but the sheer brevity of his sketch leaves room for other views. On this issue, read [**15**, **16**, **20**, and **23**].

Of course, our "completed" Newtonian demonstration is really anything but complete, since in step four, to ensure that our constructed conic motion had an inverse-square acceleration, we called on the *unproved* reversal of the Shape Theorem:

NEWTON'S ACCELERATION THEOREM. *Every conic motion about the focus has an inverse-square acceleration.*

We now intend to study the original argument for the Acceleration Theorem and then contrast the original with what we might do today, but as we return with this intention to the *Principia* (and specifically to Propositions XI, XII, and XIII in Book I), we must first page back to Proposition VI in order to understand how Newton measures orbital acceleration.

4. In May of 1686, just one month after the *Principia* was presented to the Royal Society, Halley sent news to Newton of the plans for printing and publication, but his cheerful letter ended with a sour lemon [**21**, p. 446]: "There is one thing more I ought to informe you of," he wrote,

> that M^{r} Hook has some pretensions upon the invention of y^{e} rule of the decrease of Gravity, being reciprocally as the squares of the distances from the Center. He sais you had the notion from him ... how much of this is so, you know best, as likewise what you have to do in this matter, only M^{r} Hook seems to expect you should make some mention of him, in the preface ... .

"Now is not this very fine?" sneered back Newton [**21**, p. 448],

> Mathematicians that find out, settle & do all the business must content themselves with being nothing but dry calculators & drudges & another that does nothing but pretend & grasp at all things must carry away all the invention ... And why should I record a man for an Invention who founds his claim upon an error therein & on that score gives me trouble? He imagines he obliged me by telling me his Theory, but I thought myself disobliged by being upon his own mistake corrected magisterially & taught a Theory w^{ch} every body knew & I had a truer notion of then himself.

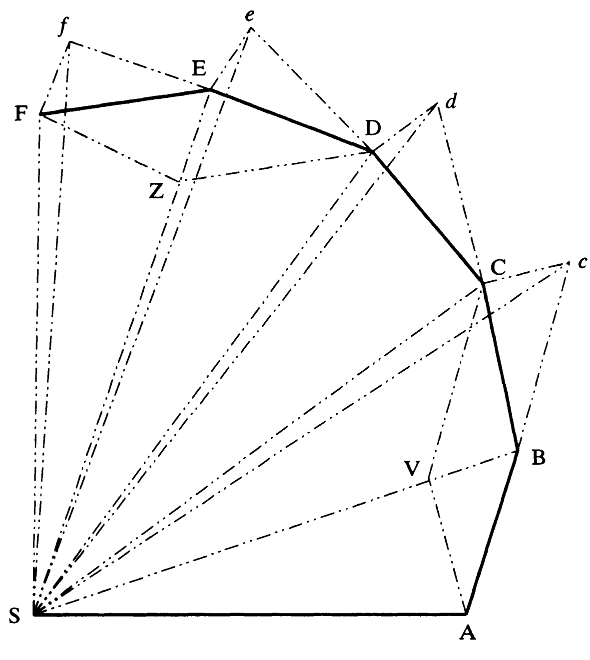
In his fury at Hooke's pretensions, Newton struck back with his pen, literally striking out almost every reference to Hooke in the entire *Principia*.

Even so, Hooke did in fact make one significant contribution to the *Principia*, for he was the first to see orbital motions as the geometric signature of a central attraction that pulls the orbiting body away from its linear inertial path. In November of 1679, as the new Secretary of the Royal Society, Hooke had asked

Newton to [22, p. 22] "please ... continue your former favors to the Society by communicating what shall occur to you that is Philosophicall," and he added,

> for my own part I shall take it as a great favor ... if you will let me know your thoughts of [my hypothesis] of compounding the celestiall motions of the planets of a direct [straight] motion by the tangent & an attractive motion towards the centrall body.

Hooke had this hypothesis as early as 1670, a time when Newton's eyes were still clouded by thoughts of "outward endeavor" and "Cartesian vortices." Still, Hooke's physical insight could take him only so far. In his hands, the hypothesis remained just that: a guess, a guess rooted in physical intuition and mechanical experiment, yet still a guess. But in Newton's hands, the hands of a soaring mathematical imagination, Hooke's hypothesis rose to an aerie of definitions, lemmas, and propositions. Look, for example, at the figure Newton draws to illustrate his proof of Propositions I and II (Section II, Book I), where we see, for the very first time,



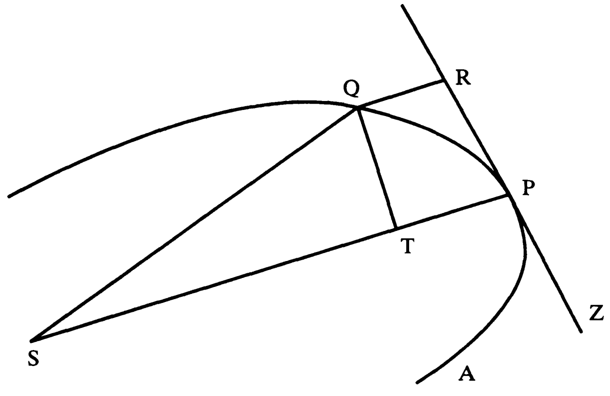
the mathematical equivalence of central attraction and the area law, and you behold, in its central attraction and deviations from the tangent, the risen form of Hooke's hypothesis.

Later, in Proposition VI, Newton fashions from Hooke's inward deviation a formula for measuring the acceleration of an orbital motion. (In the *Principia*,

accelerations for *general* motions are never even defined.) If a particle in orbital motion falls freely toward the acceleration center S, Newton may have reasoned that the particle could be thought of as instantaneously in free fall from the tangent down to its position on the orbit. In a given time t, suppose a particle moves along its orbit from P to Q. If there had been no acceleration during this time interval, the particle would have proceeded instead along the tangent at constant speed v to a location L. The deviation QL, nearly parallel to SP, would be like the "distance fallen toward S," which we would expect to be approximately $\frac{1}{2}at^2$, where a gives the acceleration at P. This suggests

$$\frac{\mathrm{QL}}{t^2} \to \frac{1}{2}a$$

as $t \to 0$. Sanding top and bottom, Newton could now have shaped the measure QL/t^2 to fit squarely into his geometric approach. First nudge L just a bit along the tangent to the position R, making the deviation QR *exactly* parallel to SP.



Because time varies as the area in orbital motions, replace t by the area of the "sector" PSQ, and the sector in turn by the approximating triangle PSQ, in the process turning t into the product $\mathrm{SP} \cdot \mathrm{QT}$—no need to keep tabs on constant factors, such as the missing 1/2 here, for Newton works with proportions, not equations—and the measure QL/t^2 into $\mathrm{QR}/(\mathrm{SP} \cdot \mathrm{QT})^2$. The limit of this ratio, as $\mathrm{Q} \to \mathrm{P}$, gauges the acceleration at P. In the *Principia*, this measure of acceleration appears as Corollary I to Proposition VI (Section II, Book I) [**11**, p. 48]. With this corollary, Newton later derives acceleration laws from orbit shapes.

> Cor I. *If a body* P *revolving about the center* S *describes a curved line* APQ, *which a right line* ZPR *touches in any point* P; *and from any other point* Q *of the curve*, QR *is drawn parallel to the distance* SP, *meeting the tangent in* R; *and* QT *is drawn perpendicular to the distance* SP; *the centripetal force will be inversely as the solid* $\mathrm{SP}^2 \cdot \mathrm{QT}^2/\mathrm{QR}$, *if the solid be taken of that magnitude which it ultimately acquires when the points* P *and* Q *coincide*.

Before we leave the topic of acceleration, we should take a moment to discuss the role of force and mass in the early sections of the *Principia*. The word 'force' appears, as it does above in Corollary I, in many of the definitions, axioms, corollaries, and propositions of the *Principia*, but in the first ten sections, where Newton attends to the one-body problem, force, and mass as well, exist literally in name only, playing no part in the *mathematics*. He may talk of 'force,' but Newton calculates accelerations. The Cartesians, Huygens and Leibniz among them, claimed that Newton, by introducing gravity, and therefore action at a distance, brought Aristotelian 'occult qualities' back into physics. But he should plead innocent to this charge. In the *Principia*'s work on orbital motions, 'force' and 'gravity' become merely convenient words, as Newton stresses the relations and laws, with no comment on *causes*. The cause of gravity comes up only in a General Scholium on the final pages of the *Principia* [**11**, p. 547]: "But hitherto I have not been able to discover the cause of those properties of gravity from phenomena," wrote Newton,

> and I frame no hypotheses; for whatever is not deduced is to be called an hypothesis; and hypotheses, whether metaphysical or physical, whether of occult qualities or mechanical, have no place in experimental philosophy. ... And to us it is enough that gravity does really exist, and act according to the laws which we have explained, and abundantly serves to account for all the motions of the celestial bodies, and of our sea.

Wouldn't Newton, that lover of geometry and curvature, have been delighted with Einstein's view that geometry, indeed the curvature of spacetime, is the very cause of gravity?

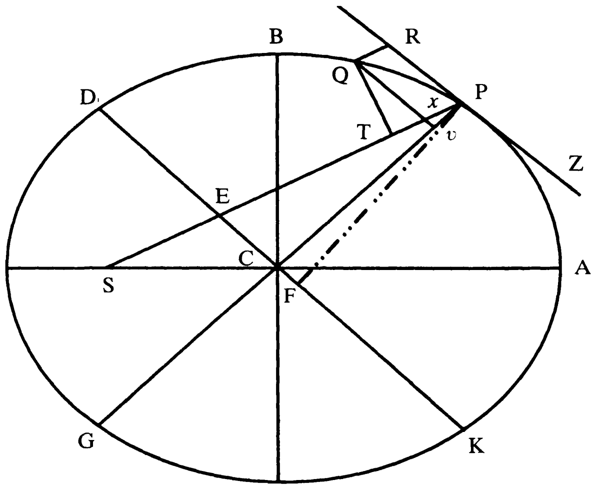
After this interlude on Newton's measure of acceleration, we remain in the past, looking for the original proof of the Acceleration Theorem in the *Principia*.

5. Wasting no time after Corollary I to Proposition VI, Newton attacks a series of problems with his new measure of acceleration. In Propositions VII through XIII, he calculates the acceleration law for circular motions about any given point, semicircular motions about a point infinitely remote, spiral motions about the pole, elliptical motions about the center, and then, in a stately section all their own, elliptical, hyperbolic, and parabolic motions about the focus. Taken together, this final triumphant trio of propositions (XI, XII, and XIII) establishes the Acceleration Theorem: *Every conic motion about the focus has an inverse-square acceleration.*

Newton could have proved the Acceleration Theorem in a single proposition covering general conic motions, but "... because of the dignity of the Problem ...," he writes, "I shall confirm the ... cases by particular demonstrations." [**11**, p. 57] These "particular demonstrations" naturally offer the same argument with minor variations, so we may safely choose one of the propositions to represent all three. Turn then to the most celebrated page of the *Principia* and to Newton's analysis for Proposition XI:

PROPOSITION XI PROBLEM VI

If a body revolves in an ellipse; it is required to find the law of the centripetal force tending to the focus of the ellipse.



In the ellipse, Newton draws conjugate diameters DK and PG, with DK parallel to the tangent RPZ. (The midpoints of parallel chords in an ellipse lie on a line, called a *diameter* of the ellipse, and the parallel chords are then called the *ordinates* of the diameter. Two diameters with the property that each bisects every chord parallel to the other are said to be *conjugate diameters*.) From Q he drops three lines: QR parallel to the focal radius SP, QT perpendicular to SP, and Qx completing the parallelogram $QxPR$. He then extends Qx until it meets PG at v and draws PF perpendicular to DK.

Newton's analysis requires the services of three lemmas, one of his own and two well known to Apollonius of Perga. (For the two Apollonian lemmas, see [**1**, I p. 15 and VII p. 31] or [**18**, p. 151 and p. 169].)

NEWTON'S LEMMA. PE = AC

LEMMA 1. *All parallelograms circumscribed about any conjugate diameters of an ellipse have equal area.*

LEMMA 2. *In an ellipse, the squares of the ordinates of any conjugate diameter are proportional to the rectangles under the segments which they make on the diameter.*

As we have seen in the previous section, Newton measures the acceleration of an orbital motion by computing the limit of the ratio

$$\frac{\mathrm{QR}}{(\mathrm{SP}\cdot\mathrm{QT})^2}$$

as $Q \to P$. To infer an inverse-square acceleration for this case of elliptical motion about the focus, he must therefore prove that $\mathrm{QR}/\mathrm{QT}^2$ has a limit independent of P. In fact, as we now show, Newton's argument reveals that $\mathrm{QT}^2/\mathrm{QR}$ tends to the *latus rectum* of the ellipse.

Because QR is Px and (by Newton's Lemma) PE is AC, the similarity of the triangles PxV and PEC implies

$$\mathrm{QR} = \frac{\mathrm{P}v \cdot \mathrm{AC}}{\mathrm{PC}}.$$

On the other hand, Newton's Lemma (again) and the similarity of the triangles QxT and PEF give

$$\mathrm{QT} = \frac{\mathrm{Q}x \cdot \mathrm{PF}}{\mathrm{AC}} = \frac{\mathrm{Q}x \cdot \mathrm{BC}}{\mathrm{CD}},$$

where the second equality follows from Lemma 1, which assures us that PF · CD = BC · AC. We infer

$$\frac{\mathrm{QT}^2}{\mathrm{QR}} = \frac{\mathrm{Q}x^2 \cdot \mathrm{BC}^2}{\mathrm{CD}^2} \cdot \frac{\mathrm{PC}}{\mathrm{P}v \cdot \mathrm{AC}} = \frac{1}{2}\mathrm{L}\frac{\mathrm{Q}x^2 \cdot \mathrm{PC}}{\mathrm{P}v \cdot \mathrm{CD}^2},$$

where we have replaced $2\mathrm{BC}^2/\mathrm{AC}$ by L. (Following Apollonius, Newton calls $2\mathrm{BC}^2/\mathrm{AC}$ the *latus rectum.*) If now Q $\to$ P, this last expression has the same limit as

$$\frac{1}{2}\mathrm{L}\frac{v\mathrm{G}}{\mathrm{PC}},$$

for Qv/Qx tends to one and Lemma 2 implies

$$\frac{\mathrm{Q}v^2}{\mathrm{P}v \cdot v\mathrm{G}} = \frac{\mathrm{CD}^2}{\mathrm{PC}^2}.$$

But vG $\to$ 2PC, so that $\frac{1}{2}\mathrm{L}(v\mathrm{G}/\mathrm{PC})$, and thus also $\mathrm{QT}^2/\mathrm{QR}$, must tend to L. This completes Newton's analysis for Proposition XI: *Every elliptical motion about the focus has an inverse-square acceleration.*

6. We have been "going under with the swirls and coming out with the eddies, following along the way the water goes," but now just one quick swirl remains: to return from the *Principia* to the present, from Newton's original work on the Acceleration Theorem to the delightful contrast of a contemporary argument.

Any conic motion $\mathbf{r} = \mathbf{r}(t)$ about the focus must satisfy the vector-conic equation (2),

$$\mathbf{r} \cdot (\mathbf{e} + \mathbf{U}) = l,$$

for some positive constant l and constant vector $\mathbf{e}$. Since $\mathbf{r}$ is an orbital motion, $\mathbf{h} = \mathbf{r} \times \mathbf{v}$ is a constant vector. Since $\mathbf{r}$ is a conic motion,

$$\mathbf{L} = \frac{l}{h^2}\mathbf{v} \times \mathbf{h} - \mathbf{U}$$

is a second constant vector (equal to the eccentricity vector $\mathbf{e}$ by (3)). Differentiating $\mathbf{L}$ yields

$$\mathbf{0} = \frac{l}{h^2}\mathbf{a} \times \mathbf{h} - \frac{\mathbf{h} \times \mathbf{r}}{r^3},$$

and taking lengths we uncover an inverse-square acceleration,

$$a = \frac{h^2}{l}\frac{1}{r^2},$$

proving again

NEWTON'S ACCELERATION THEOREM. *Every conic motion about the focus has an inverse-square acceleration.*

ACKNOWLEDGMENTS. I am indebted to J. Bruce Brackenridge, David Cook, Nathaniel Grossman, Michael Nauenberg, Alan Parks, D. T. Whiteside, Curtis Wilson, and the referees for thoughtful comments—historical, mathematical, and stylistic. In most cases, I took their advice, and where I did not, I probably should have.

REFERENCES

1. Apollonius of Perga, *Treatise on Conic Sections*, Volumes I–VII, Cambridge University Press, Cambridge, 1896.
2. J. L. Axtell, Locke, Newton, and the Two Cultures, *John Locke: Problems and Perspectives*, Cambridge University Press, Cambridge, 1969, 165–182.
3. J. B. Brackenridge, Newton's Unpublished Dynamical Principles: A Study in Simplicity, *Annals of Science* 47 (1990), 3–31.
4. J. B. Brackenridge, The Critical Role of Curvature in Newton's Developing Dynamics, *The Investigation of Difficult Things: Essays on Newton and the History of the Exact Sciences*, edited by P. M. Harman and A. E. Shapiro, Cambridge University Press, Cambridge, 1992, 231–260.
5. J. B. Brackenridge, *The Key to Newton's Dynamics: The Kepler Problem and the Principia*, University of California Press, Berkeley, 1995.
6. S. Chandrasekhar, *Newton's Principia for the Common Reader*, Oxford University Press, New York, 1995.
7. G. E. Christianson, *In the Presence of the Creator: Isaac Newton and His Times*, The Free Press, New York, 1984.
8. I. B. Cohen, *Introduction to Newton's 'Principia,'* Harvard University Press, Cambridge, 1971.
9. N. Grossman, *The Sheer Joy of Celestial Mechanics*, Birkhäuser, New York, 1995.
10. M. Nauenberg, Newton's Early Computational Method for Dynamics, *Archive for History of Exact Sciences* 46 (1994), 221–252.
11. I. Newton, *Sir Isaac Newton's Mathematical Principles of Natural Philosophy and His System of the World*, original translation by A. Motte in 1729, revised by F. Cajori, University of California Press, Berkeley, 1946.
12. I. Newton, *The Mathematical Papers of Isaac Newton*, Volumes I–VIII, edited by D. T. Whiteside, Cambridge University Press, Cambridge, 1967–1981.
13. I. Newton, *The Correspondence of Isaac Newton*, edited by A. R. Hall and L. Tilling, Cambridge University Press, Cambridge, 1975.
14. I. Newton, *The Preliminary Manuscripts for Isaac Newton's 1687 Principia 1684–1685*, introduction by D. T. Whiteside, Cambridge University Press, Cambridge, 1989.
15. B. Pourciau, On Newton's Proof That Inverse-Square Orbits Must be Conics, *Annals of Science* 48 (1991), 159–172.
16. B. Pourciau, Newton's Solution of the One-Body Problem, *Archive for History of Exact Sciences* 44 (1992), 125–146.
17. B. Pourciau, Radical Principia, *Archive for History of Exact Sciences* 44 (1992), 331–363.
18. G. Salmon, *A Treatise on Conic Sections*, Chelsea Publishing Company, New York, 1954.
19. Chuang Tzu, *Chuang Tzu: Basic Writtings*, translated by Burton Watson, Columbia University Press, New York, 1964.
20. R. Weinstock, Isaac Newton: Credit Where Credit Won't Do, *College Mathematics Journal* 25 (1994), 179–193.
21. R. Westfall, *Never at Rest: A Biography of Isaac Newton*, Cambridge University Press, Cambridge, 1980.
22. D. T. Whiteside, The Prehistory of the *Principia* From 1664 to 1686, *Notes and Records of the Royal Society of London* 45 (1991), 11–61.
23. C. Wilson, Newton's Orbit Problem: A Historian's Response, *College Mathematics Journal* 25 (1994), 193–201.

Bruce Pourciau
Department of Mathematics
Lawrence University
Appleton, Wisconsin 54912
Bruce.H.Pourciau@Lawrence.edu

Applications of Linear Algebra in Calculus

Jack W. Rogers, Jr.

1. INTRODUCTION. The concepts of basis, matrix for a linear transformation relative to bases, and change-of-basis matrix are fundamental in linear algebra, but students in an introductory class often have trouble understanding the point of applying these concepts for bases other than the standard basis for $\mathbb{R}^n$. Our object is to illustrate some applications of these concepts in solving problems with which students who have recently completed the calculus sequence should be familiar. The spaces are abstract vector spaces—finite subspaces of function spaces—not simply subspaces of $\mathbb{R}^n$; they have no obvious natural basis. We see that standard linear algebra techniques, such as matrix inversion, can be applied in place of the usual calculus techniques of substitution or integration by parts. The students may judge for themselves the relative difficulty of calculus methods *vs.* linear algebra methods—and the understanding that each provides—for these types of problems. This is not to deny the fundamental importance of substitution or integration by parts in calculus. Students are assumed to have mastered these techniques in their calculus courses and to be familiar with the problems to which they are applied. These problems can then be used to motivate new ideas in linear algebra.

2. PRELIMINARIES. We adopt the following conventions. All vectors spaces are over $\mathbb{R}$. Any sequence $\mathscr{B} = (\mathbf{b}_1, \dots, \mathbf{b}_k)$ of vectors in a vector space V determines a linear transformation $L_{\mathscr{B}}\colon \mathbb{R}^k \to V$ defined by

$$L_{\mathscr{B}}(\mathbf{x}) = L_{\mathscr{B}}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}\right) = x_1\mathbf{b}_1 + \cdots + x_k\mathbf{b}_k.$$

If $\mathscr{B} = (\mathbf{b}_1, \dots, \mathbf{b}_k)$ is a basis for V, then $L_{\mathscr{B}}$ is both injective and surjective. Thus, for every vector $\mathbf{x} \in V$ there is a unique vector $\mathbf{c} \in \mathbb{R}^k$ such that $\mathbf{x} = L_{\mathscr{B}}\mathbf{c}$. The coordinates of $\mathbf{c}$, $c_1, \dots, c_k$, are called the *coordinates of* $\mathbf{x}$ *relative to the basis* $\mathscr{B}$, and we write

$$[\mathbf{x}]_{\mathscr{B}} = \mathbf{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix}.$$

This defines the linear *coordinate transformation*, $[\cdot]_{\mathscr{B}}\colon V \to \mathbb{R}^k$, which is the inverse of $L_{\mathscr{B}}$.

3. MATRICES FOR DIFFERENTIATION. Suppose that U and V are linear spaces with bases $\mathscr{B} = (\mathbf{b}_1, \dots, \mathbf{b}_n)$ and $\mathscr{C} = (\mathbf{c}_1, \dots, \mathbf{c}_m)$, respectively, and T: $U \to V$ is linear. Suppose $\mathbf{x} = x_1\mathbf{b}_1 + \cdots + x_n\mathbf{b}_n \in U$. For $i = 1, \dots, n$, $T(\mathbf{b}_i)$ has a coordinate vector relative to $\mathscr{C}$, denoted by $[T(\mathbf{b}_i)]_{\mathscr{C}}$. These vectors form the columns of *the matrix for T relative to* $\mathscr{B}$ *and* $\mathscr{C}$, which we denote by $M_{T:\,\mathscr{C} \leftarrow \mathscr{B}}$, and which transforms the $\mathscr{B}$-coordinates of any vector in U into the $\mathscr{C}$-coordinates of

its image under T, as we see by using the linearity of T and of the coordinate transformation $[\cdot]_{\mathscr{C}}$ to obtain

$$\begin{aligned}[T(\mathbf{x})]_{\mathscr{C}} &= [x_1T(\mathbf{b}_1) + \cdots + x_nT(\mathbf{b}_n)]_{\mathscr{C}} = x_1[T(\mathbf{b}_1)]_{\mathscr{C}} + \cdots + x_n[T(\mathbf{b}_n)]_{\mathscr{C}} \\ &= [[T(\mathbf{b}_1)]_{\mathscr{C}} \cdots [T(\mathbf{b}_n)]_{\mathscr{C}}]\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = M_{T:\mathscr{C}\leftarrow\mathscr{B}}[\mathbf{x}]_{\mathscr{B}}.\end{aligned}$$

The direction of the arrow, opposite to that in $T\colon U \to V$, is chosen to preserve the order of the subscripts in the equation $[T(\mathbf{x})]_{\mathscr{C}} = M_{T:\mathscr{C}\leftarrow\mathscr{B}}[\mathbf{x}]_{\mathscr{B}}$.

3.1. An application to antidifferentiation. We consider differentiation as a linear transformation $\mathfrak{D}\colon C^1(\mathbb{R}) \to C(\mathbb{R})$ from the space of continuously differentiable functions $f\colon \mathbb{R} \to \mathbb{R}$ to the continuous functions. If $\mathscr{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ is a finite sequence of vectors in $C^1(\mathbb{R})$ such that $\mathfrak{D}$ leaves $S = \operatorname{span}\mathscr{B}$ invariant, i.e., $\mathfrak{D}(S) \subset S$, then the derivative restricted to S can be represented by an $n \times n$ matrix relative to the basis $\mathscr{B}$ alone. As a simple example, suppose $\mathscr{B} = \mathscr{C} = (\sin t, \cos t)$ and $U = V = \operatorname{span}\mathscr{B}$. Then

$$\begin{aligned}[\mathfrak{D}(\mathbf{b}_1)]_{\mathscr{C}} &= [\mathfrak{D}(\sin t)]_{\mathscr{B}} = [0\cdot\sin t + 1\cdot\cos t]_{\mathscr{B}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ [\mathfrak{D}(\mathbf{b}_2)]_{\mathscr{C}} &= [\mathfrak{D}(\cos t)]_{\mathscr{B}} = [-1\cdot\sin t + 0\cdot\cos t]_{\mathscr{B}} = \begin{bmatrix} -1 \\ 0 \end{bmatrix},\end{aligned}$$

and the matrix $D = M_{\mathfrak{D}:\mathscr{B}\leftarrow\mathscr{B}}$ for $\mathfrak{D}$ relative to $\mathscr{B}$ is $D = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. This is the matrix for a rotation of the plane 90° counterclockwise, which gives a geometric picture for the cycle of derivatives $\sin t \mapsto \cos t \mapsto -\sin t \mapsto -\cos t \mapsto \sin t$.

The matrix D is invertible, and $D^{-1} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$. Since the inverse of differentiation is integration, the columns of D^{-1} must represent antiderivatives of the basis elements:

$$\int \sin t\,dt = L_{\mathscr{B}}\left(\begin{bmatrix} 0 \\ -1 \end{bmatrix}\right) = -\cos t \quad\text{and}\quad \int \cos t\,dt = L_{\mathscr{B}}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = -\sin t.$$

Of course, there are other antiderivatives for the sine and cosine, differing from these antiderivatives by constants. They do not appear because the space of constant functions, which forms the kernel of the derivative transformation, intersects S only at the origin. Thus, relative to $\mathscr{B}$, antidifferentiation is unique.

Now consider $\int t^2e^t\,dt$. This is a typical integration by parts problem, requiring two applications to complete the integration. Instead, we look for a space containing t^2e^t that is invariant under differentiation. By successive differentiation of t^2e^t, we find one such space, $\mathscr{B} = (t^2e^t, te^t, e^t)$, and we have

$$\begin{aligned}[\mathfrak{D}(t^2e^t)]_{\mathscr{B}} &= [t^2e^t + 2te^t]_{\mathscr{B}} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} \\ [\mathfrak{D}(te^t)]_{\mathscr{B}} &= [te^t + e^t]_{\mathscr{B}} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \\ [\mathfrak{D}(e^t)]_{\mathscr{B}} &= [e^t]_{\mathscr{B}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.\end{aligned}$$

For this problem, the matrix $D = M_{\mathfrak{D}:\mathscr{B}\leftarrow\mathscr{B}}$ for $\mathfrak{D}$ relative to $\mathscr{B}$ is

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad D^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 2 & -1 & 1 \end{bmatrix}.$$

The single matrix inversion provides all the following antiderivatives, with coefficients given by the columns of D^{-1}.

$$\int t^2e^t\,dx = t^2e^t - 2te^t + 2e^t, \int te^t\,dx = te^t - e^t, \text{and} \int e^t\,dx = e^t.$$

Matrix inversion has replaced the use of integration by parts for this problem.

The same technique can be used to provide an alternative approach for many standard integration by parts problems. For $\int t\sin t\,dx$, for example, let $\mathscr{B} = (t\sin t, t\cos t, \sin t, \cos t)$; for $\int e^t \sin t\,dt$, a trickier integration by parts problem, let $\mathscr{B} = (e^t\sin t, e^t\cos t)$.

3.2. An application to linear differential equations. Suppose $f \in C^\infty$ and consider the nonhomogeneous differential equation

$$\mathscr{L}(y) = a_n y^{(n)} + \cdots + a_0 y = f. \tag{3.1}$$

This is represented in matrix form as

$$L[y]_{\mathscr{B}} = (a_nD^n + \cdots + a_0I)[y]_{\mathscr{B}} = [f]_{\mathscr{B}}.$$

Assuming $[f]_{\mathscr{B}}$ is in the image of $L(D)$, this yields the coordinates of a particular solution for (3.1). For example, consider

$$y'' + y' + y = \sin t.$$

Using the basis $\mathscr{B} = (\sin t, \cos t)$, we have $[\sin t]_{\mathscr{B}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and

$$L = D^2 + D + I = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

The solution of $L[\mathbf{y}]_{\mathscr{B}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is $[\mathbf{y}]_{\mathscr{B}} = \begin{bmatrix} 0 \\ -1 \end{bmatrix} = [-\cos \mathrm{t}]_{\mathscr{B}}$, yielding $y = -\cos t$ as a particular solution for the differential equation.

This is related to the method of undetermined coefficients for finding a particular solution [2]; using the matrix for the derivative simplifies the computations.

4. CHANGE-OF-BASIS AND CERTAIN TRIGONOMETRIC INTEGRALS. We turn now to an application for the change-of-basis matrix.

Suppose $\mathscr{B}$ and $\mathscr{C}$ are two bases for the same k-dimensional vector space, V. The change from $\mathscr{B}$-coordinates to $\mathscr{C}$-coordinates is a linear transformation, with an associated matrix $P_{\mathscr{C}\leftarrow\mathscr{B}}$ satisfying

$$[\mathbf{x}]_{\mathscr{C}} = P_{\mathscr{C}\leftarrow\mathscr{B}}[\mathbf{x}]_{\mathscr{B}}.$$

It is the matrix for the identity transformation relative to the bases $\mathscr{B}$ and $\mathscr{C}$:

$$P_{\mathscr{C}\leftarrow\mathscr{B}} = M_{I:\mathscr{C}\leftarrow\mathscr{B}} = \left[[\mathbf{b}_1]_{\mathscr{C}} \cdots [\mathbf{b}_k]_{\mathscr{C}}\right].$$

4.1. Integrating powers of $\cos x$. The integration of $\cos^n t$ is usually accomplished at the calculus level by substitution if n is odd, and by application of the double-angle formula $\cos 2t = 2\cos^2 t - 1$, repeated as necessary, if n is even. An alternative, considering this as a change-of-basis problem in linear algebra, provides a unified approach to integrating any polynomial in $\cos t$.

Let $\mathscr{B}_{n+1} = (1, \cos t, \dots, \cos nt)$. There are several ways to show that these functions are independent; here, we simply make that assumption. Thus $\mathscr{B}_{n+1}$ is a basis for the span, S_{n+1}, of its terms.

Adding $\cos(k+1)t$ and $\cos(k-1)t$ and regrouping yields the recursion

$$\cos(k+1)t = 2\cos kt \cos t - \cos(k-1)t.$$

For the first few values of k, we have

$$\begin{aligned}
\cos 0t &= 1\\
\cos 1t &= \cos t\\
\cos 2t &= 2(\cos 1t)\cos t - \cos 0t = -1 + 2\cos^2 t\\
\cos 3t &= 2(\cos 2t)\cos t - \cos t = 2(-1 + 2\cos^2 t)\cos t - \cos t\\
&= -3\cos t + 4\cos^3 t\\
\cos 4t &= 1 - 8\cos^2 t + 8\cos^4 t\\
\cos 5t &= 5\cos t - 20\cos^3 t + 16\cos^5 t\\
&\ \vdots
\end{aligned} \tag{4.2}$$

In general, this recursion expresses elements of $\mathscr{B}_{n+1}$ as *Chebyshev polynomials* (see Section 5.1) in $\cos t$, i.e., as linear combinations of the elements of $\mathscr{C}_{n+1} = (1, \cos t, \dots, \cos^n t)$. Since $\mathscr{C}_{n+1}$ is a sequence of $n+1$ vectors spanning the $(n+1)$-dimensional space $S_{n+1}, \mathscr{C}_{n+1}$ is also a basis for S_{n+1}. The equations (4.2) show how to form the change-of-basis matrix $P_{\mathscr{C} \leftarrow \mathscr{B}}$ (we suppress subscripts and superscripts for the bases unless needed for clarity). The inverse change-of-basis matrix, $P_{\mathscr{B} \leftarrow \mathscr{C}}$, converts a polynomial in $\cos t$, which is difficult to integrate, into a linear combination of terms of the form $\cos nt$, which is easy to integrate.

For example, for $n = 2$, equations (4.2) show that

$$P_{\mathscr{C} \leftarrow \mathscr{B}} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \quad \text{and} \quad P_{\mathscr{B} \leftarrow \mathscr{C}} = P^{-1}_{\mathscr{C} \leftarrow \mathscr{B}} = \begin{bmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}.$$

Thus, for an arbitrary second-degree polynomial p in $\cos t$, we have

$$\begin{aligned}
[p]_{\mathscr{C}} &= \left[c_0 + c_1 \cos t + c_2 \cos^2 t\right]_{\mathscr{C}} = \begin{bmatrix} c_0 & c_1 & c_2 \end{bmatrix}^T\\
&\Rightarrow [p]_{\mathscr{B}} = P_{\mathscr{B} \leftarrow \mathscr{C}}[p]_{\mathscr{C}} = \begin{bmatrix} c_0 + \frac{1}{2}c_2 & c_1 & \frac{1}{2}c_2 \end{bmatrix}^T\\
&\Rightarrow p = \left(c_0 + \tfrac{1}{2}c_2\right) + c_1 \cos t + \tfrac{1}{2}c_2 \cos 2t,
\end{aligned}$$

which is easily integrated.

4.2. The case for even n. We have seen that polynomials in $\cos t$ can be integrated using an appropriate change of basis. Integrating $\cos^n t$ for odd n can be done by substitution, so we can focus our attention on the even powers. This problem can also be solved using a change-of-basis matrix—one that is half the size of the one we must deal with if we include the odd powers.

Adding the expressions for $\cos(k+2)t$ and $\cos(k-2)t$ yields, for $k > 1$,

$$\begin{aligned}
\cos(k+2)t &= 2(\cos kt)(\cos 2t) - \cos(k-2)t\\
&= 2(\cos kt)(-1 + 2\cos^2 t) - \cos(k-2)t.
\end{aligned}$$

Thus,

$$\begin{aligned}
\cos 0t &= 1\\
\cos 2t &= -1 + 2\cos^2 t\\
\cos 4t &= 2(\cos 2t)(-1 + 2\cos^2 t) - \cos 0t = 2(-1 + 2\cos^2 t)^2 - 1\\
&= 1 - 8\cos^2 t + 8\cos^4 t\\
\cos 6t &= 2(1 - 8\cos^2 t + 8\cos^4 t)(-1 + 2\cos^2 t) - (-1 + 2\cos^2 t)\\
&= -1 + 18\cos^2 t - 48\cos^4 t + 32\cos^6 t\\
&\ \vdots
\end{aligned} \tag{4.3}$$

Continued application of the recursion formula yields, for every k, an expression for $\cos 2kt$ as a polynomial in *even* powers of $\cos t$. Consequently, the sequences $\mathscr{C}_{n+1}^{(e)} = (1, \cos^2 t, \dots, \cos^{2n} t)$ and $\mathscr{B}_{n+1}^{(e)} = (1, \cos 2t, \dots, \cos 2nt)$ are bases for the same $(n+1)$-dimensional subspace of S_{2n+1}, and the change-of-basis matrix provides an organized method for integrating even polynomials in $\cos x$, including, as a special case, $\cos^{2n} t$.

For example, for $\int \cos^6 t\, dt$, $n = 3$, and we have

$$P_{\mathscr{C} \leftarrow \mathscr{B}} = \begin{bmatrix} 1 & -1 & 1 & -1\\ 0 & 2 & -8 & 18\\ 0 & 0 & 8 & -48\\ 0 & 0 & 0 & 32 \end{bmatrix}$$

and

$$P_{\mathscr{B} \leftarrow \mathscr{C}} = P_{\mathscr{C} \leftarrow \mathscr{B}}^{-1} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{3}{8} & \frac{5}{16}\\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{15}{32}\\ 0 & 0 & \frac{1}{8} & \frac{3}{16}\\ 0 & 0 & 0 & \frac{1}{32} \end{bmatrix}.$$

The coefficients of $\cos^6 t$ are in the last column of this matrix, $\cos^6 t = \frac{5}{16} + \frac{15}{32}\cos 2t + \frac{3}{16}\cos 4t + \frac{1}{32}\cos 6t$, from which the integral is easily obtained. The formula used here is a typical three-term recursion, i.e., except for $\cos t$, which does not depend on k, the value of $\cos(k+2)t$ depends only on two earlier even values, $\cos kt$ and $\cos(k-2)t$.

For comparison, the standard calculus approach uses the double-angle formula $\cos^{2k} t = (\cos^2 t)^k = [\frac{1}{2}(1 + \cos 2t)]^k$, integrating constants and odd powers by substitution as they occur, until k is reduced to 1 or 0. More formally, suppose all even powers of $\cos^{2i} t$, for $i < k$, have been expressed in terms of odd powers of $\cos jt$. Then the binomial formula yields

$$\cos^{2k} t = \frac{1}{4}[(1 + \cos 2t)]^k = \frac{1}{4}\sum_{i=0}^{k} \binom{k}{i} \cos^i 2t,$$

and the even powers in this expression can be replaced by the expressions already calculated. For $n = 6$, this algorithm yields

$$\cos^6 t = \tfrac{5}{16} + \tfrac{3}{8}\cos 2t + \tfrac{3}{16}\cos 4t + \tfrac{1}{8}\cos^3 2t.$$

This requires all the earlier even powers of $\cos jt$, not just the previous two, and generates a more complicated basis; it does, however, avoid inversion.

4.3. Integrating powers of $\sin x$. For powers of the sine function, we let $\mathscr{E}_{n+1} = (1, \sin t, \dots, \sin^n t)$, and take $\mathscr{D}_{n+1} = (1, \sin t, \cos 2t, \dots, \cos nt)$ if n is even, and $\mathscr{D}_{n+1} = (1, \sin t, \cos 2t, \dots, \sin nt)$ if n is odd. We need two recursion formulas.

Subtracting $\cos(k-1)t$ from $\cos(k+1)t$ yields

$$\cos(k+1)t = -2\sin kt \sin t + \cos(k-1)t.$$

Subtracting $\sin(k-1)t$ from $\sin(k+1)t$ yields

$$\sin(k+1)t = 2\cos kt \sin t + \sin(k-1)t.$$

Using the first equation for odd k and the second for even k, we obtain, by induction, expressions for the elements of $\mathscr{D}_{n+1}$ in terms of those in $\mathscr{E}_{n+1}$, and hence the components of $P_{\mathscr{D} \leftarrow \mathscr{E}}$. The first few are

$$\begin{aligned}
\cos 0t &= 1 \\
\sin 1t &= \sin t \\
\cos 2t &= \cos 0t - 2\sin t \sin t = 1 - 2\sin^2 t \\
\sin 3t &= 2\cos 2t \sin t + \sin t = 2(1 - 2\sin^2 t)\sin t + \sin t = 3\sin t - 4\sin^3 t \\
\cos 4t &= 1 - 8\sin^2 t + 8\sin^4 t \\
\sin 5t &= 5\sin t - 20\sin^3 t + 16\sin^5 t \\
&\vdots
\end{aligned} \tag{4.4}$$

There is an obvious similarity between these formulas and those for the powers of the cosine, which can be explained using the basic definition of the Chebyshev polynomials (see Section 5.1).

For $n = 2$, these equations show that

$$P_{\mathscr{E} \leftarrow \mathscr{B}} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \quad \text{and} \quad P_{\mathscr{B} \leftarrow \mathscr{E}} = P^{-1}_{\mathscr{E} \leftarrow \mathscr{B}} = \begin{bmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{2} \end{bmatrix}.$$

Thus, for an arbitrary second-degree polynomial p in $\sin t$, we have

$$\begin{aligned}
[p]_{\mathscr{E}} &= \left[c_0 + c_1 \sin t + c_2 \sin^2 t\right]_{\mathscr{E}} = \begin{bmatrix} c_0 & c_1 & c_2 \end{bmatrix}^T \\
&\Rightarrow [p]_{\mathscr{B}} = P_{\mathscr{B} \leftarrow \mathscr{E}}[p]_{\mathscr{E}} = \begin{bmatrix} c_0 + \frac{1}{2}c_2 & c_1 & -\frac{1}{2}c_2 \end{bmatrix}^T \\
&\Rightarrow p = \left(c_0 + \tfrac{1}{2}c_2\right) + c_1 \sin t - \tfrac{1}{2}c_2 \cos 2t,
\end{aligned}$$

which is easily integrated.

As these formulas indicate, it follows by induction that $(1, \sin^2 t, \ldots, \sin^{2n} t)$ and $(1, \cos 2t, \ldots, \cos 2nt)$ span the same space, as do $(\sin t, \ldots, \sin^{2n-1} t)$ and $(\sin t, \ldots, \sin(2n+1)t)$.

5. CONNECTIONS. The ideas discussed so far are related to some other topics, which we briefly discuss in this section.

5.1. Chebyshev polynomials. As indicated by (4.3), $\cos nt$ can be expressed as a polynomial in $\cos t$ for each $n \geq 0$. The polynomial itself is defined as $T_n(x) = \cos(n \arccos x)$ for $x \in [-1, 1]$, so that, for $t \in [0, \pi]$, $T_n(\cos t) = \cos(n \arccos \cos t) = \cos nt$. These polynomials, called *Chebyshev polynomials*, find application in approximation theory [1]. To see a relationship between the polynomials in (4.3) and those in (4.4), we consider, for $t \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and $n \geq 0$,

$$T_n(\sin t) = \cos(n \arccos \sin t) = \cos\left(\frac{n\pi}{2} - nt\right),$$

yielding the following equations for $k \geq 0$, and explaining the relationship.

$$T_{2k}(\sin t) = (-1)^k \cos 2kt \qquad T_{2k+1}(\sin t) = (-1)^k \sin(2k+1)t$$

5.2. Finite Fourier series. The Fourier series of order n is

$$a_0 + \sum_{i=1}^{n} a_i \sin it + \sum_{i=1}^{n} b_i \cos it,$$

and integrating such an expression is straightforward. The approach in Section 4 can be thought of as expressing polynomials in sin, or polynomials in cos, as finite Fourier series, and then integrating. It is in fact true that any polynomial in sin and cos, $\sum_{i,j=0}^{n} a_{ij} \sin^i t \cos^j t$, has a finite Fourier series. It is sufficient to consider $\sin^i t \cos^j t$. If i is even, the identity $\sin^2 t + \cos^2 t = 1$ can be used to convert this to a polynomial in $\cos t$, which we have already seen to have a finite Fourier series. If i is odd, the same identity converts $\sin^i t \cos^j t$ to the form $(\sin t)p(\cos t)$, where p is a polynomial. Hence, it suffices to show that $\sin t \cos^n t$ has a finite Fourier series for each n.

Adding $\sin(k-1)t$ and $\sin(k+1)t$ yields the recursion

$$\sin(n+1)t = 2 \sin nt \cos t - \sin(n-1)t,$$

and clearly $\sin 0t$ and $\sin 1t$ are of the form $(\sin t)p(\cos t)$. Applying the recursion and induction, we see that $\sin nt$ can be expressed in this form for all n. The first few expressions are

$$\begin{aligned}
\sin 0t &= 0 \\
\sin 1t &= \sin t \\
\sin 2t &= 2(\sin t)\cos t - (0) = 2 \sin t \cos t \\
\sin 3t &= 2(2 \sin t \cos t)\cos t - (\sin t) = \sin t(-1 + 4\cos^2 t) \\
\sin 4t &= \sin t(-4 \cos t + 8 \cos^3 t) \\
\sin 5t &= \sin t(-1 - 12\cos^2 t + 16 \cos^4 t) \\
&\vdots
\end{aligned}$$

Thus, $(\sin t, \dots, \sin nt)$ and $(\sin t, \sin t \cos t, \dots, \sin t \cos^n t)$ are bases for the same space, so that $\sin t \cos^n t$ can be expressed as a finite Fourier series using only sines, which can be obtained, as before, by matrix inversion. This completes the argument.

REFERENCES

1. R. L. Burden and J. D. Faires, *Numerical Analysis*, PWS-Kent, Boston, fourth ed., 1989.
2. D. A. Sanchez, R. C. Allen, Jr. and W. T. Kyner, *Differential Equations*, Addison-Wesley, Reading, second ed., 1988.

Mathematics Department
Auburn University, Alabama 36849
jrogers@mail.auburn.edu

Periodicity, Quasiperiodicity, and Bieberbach's Theorem on Crystallographic Groups

A. Vince

1. INTRODUCTION. This article contains an elementary proof of a fundamental geometric theorem of Bieberbach. Moreover, it affords the opportunity to digress onto subjects that motivated the proof—periodicity and quasiperiodicity. The proof is in Sections 5 and 6. Most of the article consists of observations on isometries of Euclidean space (Section 2), crystallographic groups (Section 3), and the role of Bieberbach's theorem in the theory of crystals and quasicrystals (Section 4).

A *crystallographic group* is a discrete, cocompact group of isometries of n-dimensional Euclidean space. All terms in this definition are explained in Section 3. For now, it suffices to say that the two-dimensional crystallographic groups, often called wallpaper groups, are familiar as symmetry groups of tilings of the plane, and the three-dimensional groups arise as symmetry groups of crystals. There are exactly 2 one-dimensional, 17 two-dimensional, and 230 three-dimensional crystallographic groups. In dimension four there are 4,783 crystallographic groups [2]; this enumeration relys heavily on the computer. The exact number in higher dimensions is unknown. The eighteenth of Hilbert's famous problems posed at the 1900 International Congress of Mathematicians asks, in part, whether the number of crystallographic groups is finite in all dimensions. An affirmative answer was provided by Bieberbach [1] in papers that appeared in 1911 and 1912. The two- and three-dimensional crystallographic groups were first classified in the 1890's by Fedorov [7] and, independently, by Schoenflies [14]. The classification of the three-dimensional crystallographic groups can be found in many texts on mathematical crystallography, but these texts usually assume the following result. This same result is the main step in Bieberbach's solution of Hilbert's eighteenth problem.

Theorem 1 (Bieberbach). *If G is an n-dimensional crystallographic group, then G contains translations in n linearly independent directions.*

Bieberbach's proof of Theorem 1 [1] depends on a nontrivial number theoretic result concerning the approximation of irrational numbers by rationals. More recent treatments by Wolf [17] and Charlap [5], also somewhat technical, are based on a proof of Frobenius [8] that appeared shortly after Bieberbach's proof. A shorter proof by P. Buser [4] was a result of his study of Gromov's work on almost flat manifolds. Gromov, in turn, has stated that his work on almost flat manifolds resulted from an attempt to understand the Bieberbach theorem [5]. Our proof is intended to be accessible to anyone with a basic undergraduate knowledge of abstract and linear algebra.

The concept that plays the central role in the proof is what we call the *axis* of an isometry g, the largest subspace of $\mathbb{R}^n$ on which g acts as a pure translation. A main step in the proof, a result also proved by Buser **[4]**, is an analog in $\mathbb{R}^n$ of the well known Crystallographic Restriction in $\mathbb{R}^2$ and $\mathbb{R}^3$.

2. ISOMETRIES. An *isometry* is a mapping of $\mathbb{R}^n$ onto itself that preserves distance. The following representation of an isometry is well known and very easy to prove given the fact that an isometry with a fixed point is an orthogonal transformation. Given any point $p \in \mathbb{R}^n$, an isometry g can be expressed as the composition of an orthogonal transformation A, centered at p, and a translation:

$$g(x) = Ax + a. \tag{2.1}$$

The orthogonal map A will be referred to as the *rotational part* and translation by a the *translational part* of g. The rotational part is, up to conjugacy, independent of the point p. The main result in this section is a refinement of (2.1), obtained by making an appropriate choice of the origin p.

Lemma 1. *Let g be an isometry of $\mathbb{R}^n$. There exists a unique affine subspace F satisfying the following properties: (a) g is a translation when restricted to F, and (b) F is maximal with respect to property (a). Moreover, if the origin is chosen to lie in F, then*

$$g(x) = Qx + q,$$

where Q is orthogonal, F is the set of fixed points of Q, and $q \in F$.

The subspace F of Lemma 1 will be called the *axis* of g and denoted *axis*(g). As an application of Lemma 1, consider any isometry g of $\mathbb{R}^3$. If *axis*$(g) = \mathbb{R}^3$, then, according to Lemma 1, g is a translation. If *axis*(g) is a plane π, then g is the composition of a reflection through π and a translation in a direction along π. Such an isometry is called a *glide reflection* (or a *reflection* if the translation is the identity). If *axis*(g) is a line l, then g is the composition of a non-identity rotation and a translation along l. Such an isometry is called a *screw displacement* (or a *rotation* if the translation is the identity). Finally, if *axis*(g) is a point p, then g is an orthogonal transformation having only p as fixed point. Such an orthogonal transformation has canonical form

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & -1 \end{pmatrix},$$

which is the composition of a non-identity rotation about a line l with a reflection in a plane perpendicular to l. Such an isometry is called a *rotary reflection*. Thus Lemma 1 provides the following classification: every 3-dimensional isometry is a translation, rotation, reflection, glide, screw or rotary reflection.

Proof of Lemma. As in (2.1), write $g(x) = Ax + a$. Let V be the subspace of fixed points of A and $V^\perp$ the orthogonal complement of V. Note that both V and $V^\perp$ are invariant under A. Let q and $q^\perp$ be the components of a in the subspaces V and $V^\perp$, respectively. Since $I - A$ is nonsingular when restricted to $V^\perp$, the affine subspace $F = (I - A)^{-1}q^\perp$ is not empty. For $x \in F$ we have $Ax = x - q^\perp$, which implies that $g(x) = Ax + a = x + (a - q^\perp) = x + q \in F$. Therefore g is a translation when restricted to F. Define $Qx = Ax + q^\perp$. Then F is the set of fixed

points of Q; Q is orthogonal because it has a fixed point; and $g(x) = Qx + q$. We leave to the reader the routine exercise of showing that F is unique, i.e., that there does not exist even a one dimensional subspace, not contained in F, upon which G acts as a translation. ■

3. CRYSTALLOGRAPHIC GROUPS, DELAUNAY SETS AND VORONOI TILINGS. An n-dimensional *crystallographic group* G is a discrete, cocompact subgroup of isometries of $\mathbb{R}^n$. *Discrete* means that any ball contains at most finitely many points in the G-orbit of any point. *Cocompact* means that the quotient space $\mathbb{R}^n/G$ is compact, where the quotient is the set of orbits with the quotient topology. A less abstract, but equivalent, definition of crystallographic group is more appropriate for our purpose. A set X of points of $\mathbb{R}^n$ is called an (r, R)-*Delaunay set*, or simply Delaunay set, if

(1) X is *discrete*: there is a number r such that every ball of radius r centered at a point of X contains no other points of X.
(2) X is *uniform*: there is a number R such that every ball of radius R contains a point of X.

Let G be a group of isometries of $\mathbb{R}^n$ and p any point of $\mathbb{R}^n$. Then G is a crystallographic group if and only if the orbit of p is a Delaunay set. This can be restated in terms of Voronoi tilings as follows. Let P be the orbit of any point under the action of a group G of isometries of $\mathbb{R}^n$. For any $p \in P$, let D_p denote the *Voronoi region* of p. This is the set of points at least as close to p as to any other point of P:

$$D_p = \{x \in \mathbb{R}^n \colon |x - p| \leq |x - y| \text{ for all } y \in P\}.$$

The Voronoi region D_p is the intersection of half space determined by the perpendicular bisectors of the line segments joining p to each of the other points of P. The group G is a crystallographic group if and only if each Voronoi region $\{D_p | p \in P\}$ is a bounded convex polytope. In particular, the Voronoi regions of any orbit of a crystallographic group tile $\mathbb{R}^n$; all the tiles are congruent; G acts transitively on these tiles; and the action of $g \in G$ on a single tile completely determines g.

This definition makes it easy to prove a first approximation to Bieberbach's theorem, a result Buser [4] calls Mini-Bieberbach. It states that an n-dimensional crystallographic group must contain n isometries that are nearly translations, in the sense that the translational parts are linearly independent and the rotational parts are close to the identity. As a measure of the proximity of the rotational part of an isometry g to the identity, define

$$rot(g) = \max_{x \in \mathbb{R}^n} \frac{|Ax - x|}{|x|},$$

where A is the rotational part of g.

Lemma 2 (Mini-Bieberbach). *Let G be an n-dimensional crystallographic group. Given any point p and any $\varepsilon > 0$, there exist in G elements $g_i(x) = Ax + a_i$, $i = 1, 2, \ldots, n$, centered at p, such that*

(1) $rot(g_i) < \varepsilon$ *for all i* *and* (2) $\{a_1, a_2, \ldots, a_n\}$ *is linearly independent.*

Proof: Consider the Voronoi tiling with respect to the orbit of the point p. Further, let b be an arbitrary direction and consider the sequence $\{D_i\}$ of tiles that

intersect the ray with endpoint p and direction b. Let $g_i(x) = A_i x + a_i$ be an element of G with rotational part A_i centered at p and such that g_i takes D_0 to D_i, where D_0 is the tile centered at p. Because the orthogonal group is compact, $\{g_i\}$ has a convergent subsequence. For us this means that there must exist two isometries g_j and g_k with the properties: (1) A_j and A_k are sufficiently close in the sense that $rot(g_k \circ g_j^{-1}) < \varepsilon$, and (2) $|a_j - a_k|$ is sufficiently large so that the angle between b and the vector from $g_j(p)$ to $g_k(p)$ is less than ε. Then the element $g = g_k \circ g_j^{-1}$ satisfies statement (1) in the lemma. The lemma follows by repeating this argument where, at the k^{th} stage, b is chosen orthogonal to the subspace spanned by $a_1, a_2, \dots, a_{k-1}$. ■

4. CRYSTALS AND QUASICRYSTALS. With the atoms and molecules of a real crystal in mind, define a *crystal* as the image of a finite number of points of $\mathbb{R}^n$ under a group generated by n linearly independent translations. The *symmetry group*, $\mathbf{sym}(X)$, of a crystal X is the group of isometries that leave the points of X as a whole invariant. A crystal clearly has the following properties: X is discrete; X is *periodic*, which means that $\mathbf{sym}(X)$ contains translations in n linearly independent directions; X is the union of finitely many lattices, a *lattice* being the image of a single point under a group generated by n linearly independent translations.

The notions, crystal and crystallographic group, are intimately related, as described in Theorem 2. Although it would be surprising if it were otherwise, this theorem is illustrative for a couple of reasons. First, it is another consequence of Bieberbach's theorem. Second, the proof uses two essential ingredients in crystallographic analysis, the translation group and the point group. Let g be an element of a crystallographic group G and let $p \in \mathbb{R}^n$. Consider the representation (2.1): $g(x) = Ax + a$, with respect to p. The mapping ϕ: $g \mapsto A$ induces a homomorphism of G into the orthogonal group. The kernel of ϕ is the *translation subgroup* T of G; the image of ϕ is the *point group* at p.

Theorem 2. *A set X of points in Euclidean space is a crystal if and only if X is discrete and* $\mathbf{sym}(X)$ *is a crystallographic group.*

Proof: Assume that X is a crystal, G its symmetry group, and T the subgroup of G generated by the n independent translations that define X. If $p \in X$ then $G(p)$, the orbit of p, is the union of finitely many lattices since $G(p)$ is invariant under T. Therefore, $G(p)$ is a Delaunay set, so G is a crystallographic group.

In the other direction, let G be the symmetry group of X, and let T be the translation subgroup of G, which, by Bieberbach's theorem, is generated by n independent translations. Let $p \in X$ and let L be the lattice that is the image of p under the action of T. From the fact that T is normal in G, it is easy to show that L is invariant under the action of the point group Φ at p and that, for any $A \in \Phi$, A is completely determined by its action on finitely many points of L. Then Φ, being a group of permutations of these points, is finite. Being isomorphic to Φ, the quotient G/T is also finite. Express $G = \bigcup_i Tg_i$ as the disjoint union of finitely many cosets of T. Denoting by D_p the Voronoi region at p, there are finitely many points in $X_p = X \cap D_p$ because X is discrete. Now we have $X = G(X_p) = (\bigcup_i Tg_i)(X_p) = T[\bigcup_i g_i(X_p)]$. Therefore X is the image of finitely many points under the action of the translation subgroup T. By Theorem 1, T contains translations in n linearly independent directions, so, by definition, X is a crystal. ■

It was actually quasicrystals, rather than crystals, that drew our attention to Bieberbach's result. A decade ago Shechtman, Blech, Gratias, and Cahn **[16]** discovered the first "quasicrystal," an alloy of aluminum and managanese whose electron diffraction pattern consisted of sharp spots exhibiting a 5-fold symmetry. This elicited great excitement in solid-state science for the following reason. A distinct diffraction pattern with sharp spots, called Bragg peaks, is evidence of "long range order," which, until that time, meant a crystal structure. On the other hand, the well known *Crystallographic Restriction* states that the only rotational symmetry possible for a crystal in two or three dimensions is 2, 3, 4, or 6-fold symmetry. In other words, a crystal structure and the observed 5-fold symmetry are incompatible. Since this original discovery, various similar materials (aluminum-lithium-copper, uranium-palladium-silicon, and other compositions) have been discovered and analyzed, and the consensus among solid-state scientists is that these materials cannot be explained within the framework of a periodic structure, that they are truly new. The "long range order" in quasicrystals, whatever is causing the Bragg peaks in the electron diffraction, is often referred to as "quasiperiodicity."

In any study of quasiperiodicity, a minimum that should be required of a set X of points is that X be a Delaunay set. However, this alone implies little about global order, an example being the molecules of a gas in a closed container. Senechal and Taylor **[15]** inquire about the consequences of requiring the following additional local congruence property. For $x \in X$ and real number ρ, let $N_\rho(x)$ denote the intersection of X with the ball of radius ρ centered at x.

Property $\mathbf{N}_\rho$: For any two points $x, y \in X$, the neighborhoods $N_\rho(x)$ and $N_\rho(y)$ are congruent by a congruence taking x to y.

Unfortunately, as Senechal and Taylor point out, a theory based on the local regularity Property $\mathbf{N}_\rho$ will not be interesting because it already implies that X is a crystal.

Theorem 3. *Let X be an (r, R)-Delaunay set in $\mathbb{R}^n$. There exists a number ρ, depending only on r, R, and n, such that if property $\mathbf{N}_\rho$ holds, then X is a crystal.*

Theorem 3 is again a consequence of Bieberbach's theorem. The proof of Theorem 3 is in two parts. First, in 1976 Delaunay and his colleagues **[6]** gave an elegant proof that, under the conditions of Theorem 3, the symmetry group G of X acts transitively. Since the orbit X of G is a Delaunay set, G is a crystallographic group. Theorem 3 now follows directly from Theorem 2, which, in turn, was a consequence of Bieberbach's theorem.

Theorem 3 implies that any investigation into quasiperiodicity requires ideas more subtle than the local homogeneity given by property $\mathbf{N}_\rho$. Advances in this direction have been made by Penrose **[12]**, de Bruijn **[3]**, Kramer and Neri **[10]**, Katz and Duneau **[9]**, Mozes **[11]**, Radin **[13]**, and many others, but these results lie outside the scope of this note.

5. CONJUGACY IN A CRYSTALLOGRAPHIC GROUP. Very informally, to say that two isometries of Euclidean space are *conjugate* means that they do the same thing, but in different places. In $\mathbb{R}^3$, for example, the conjugate kgk^{-1} of a $(\pi/2)$-rotation g about a line l is a $(\pi/2)$-rotation about the image line $k(l)$. Lemma 3 is a more formal statement. The notation is as follows. Let g be an isometry; use Lemma 1 to express it in the form $g(x) = Qx + q$, where Q is

orthogonal and $q \in axis(g)$. Define $trans(g) = q$, where $trans(g)$ is considered as a free vector so, in statement (2) of Lemma 3, k maps both the initial and terminal point of the vector.

Lemma 3. *If g and k are isometries and $h = kgk^{-1}$ then*

$$(1)\ axis(h) = k(axis(g)) \qquad (3)\ rot(h) = rot(g)$$
$$(2)\ trans(h) = k(trans(g)) \quad (4)\ rot(hg^{-1}) \leq 2rot(k)rot(g).$$

Proof: The first three statements are routine to verify. The following proof of statement (4) is due to Buser [4]. Let A and B be the orthogonal parts of g and k, respectively, centered at the same point. Then $BAB^{-1}A^{1} - I = ((B-I)(A-I) - (A-I)(B-I))B^{-1}A^{-1}$ and, since $|B^{-1}A^{-1}x| = |x|$, it follows that

$$rot(kgk^{-1}g^{-1}) = rot(BAB^{-1}A^{-1}) \leq 2rot(B)rot(A) = 2rot(k)rot(g). \qquad \blacksquare$$

Although somewhat technical, the next lemma is essential to the proof of Bieberbach's Theorem. A rough sketch of how it comes into play is as follows. Let G be a crystallographic group. Mini-Bieberbach (Lemma 2) implies the existence of n isometries with translational parts in independent directions and with rotational parts that are close to the identity. To prove Bieberbach's theorem it remains to show only that each such isometry g must necessarily be a translation. Assume the contrary, that g is not a translation. Under this assumption, a certain set C of conjugates of g, each distinct from g, is not empty. Lemma 4 is used to prove that $axis(\bar{g})$ and $axis(g)$ are not too close to each other if $\bar{g} \in C$. So among the isometries in C, let h have axis closest to the axis of g. Then it can be shown that $axis(hgh^{-1})$ is even closer to $axis(g)$ than is $axis(h)$, a contradiction if $hgh^{-1} \in C$. Lemma 4 is required again to show that $hgh^{-1} \in C$. The complete proof appears in Section 6.

As apparent from this outline, the minimum distance between the axes of two isometries g and h plays a crucial role. We use the notation

$$d(g,h) = \min\{|x-y| : x \in axis(g),\ y \in axis(h)\}.$$

Lemma 4. *If g is an element of a crystallographic group, then there exist positive numbers δ and c with the following property. Let $h = kgk^{-1}$ be a conjugate of g. If $rot(k) < \delta$ and either*

$$(1)\quad d(g,h) \leq c \qquad \text{or} \qquad (2)\quad hgh^{-1} = g,$$

then $h = g$.

Proof: Let p and $\bar{p}$ be closest points on $axis(g)$ and $axis(h)$, respectively, and consider the Voronoi tiling with respect to the orbit of p. Choose c small enough so that, if $d(g,h) \leq c$, then $\bar{p}$ lies in the interior of tile D_p. Choose $\delta < \sqrt{2}$ and, in addition, small enough so that both of the following conditions are satisfied.

(1) If $\bar{p}$ lies in the interior of tile D_p and $rot(k) < \delta$, then $g(D_p) \cap h(D_p) \neq \varnothing$. This is possible due to statement (2) of Lemma 3.
(2) If $f(D_p) = D_p$ and $rot(f) < 4\delta$, then f must act as the identity on D_p. This is possible because D_p is a bounded polytope with finite symmetry group.

Now assume that $rot(k) < \delta$ and $d(g,h) \leq c$. By statement (1) we have $g(D_p) \cap h(D_p) \neq \varnothing$, which implies that $g(D_p) = h(D_p)$ and $g^{-1}h(D_p) = D_p$. By parts (2)

and (4) of Lemma 3, $rot(g^{-1}h) = rot(hg^{-1}) \leq 2rot(k)rot(g) \leq 4rot(k) \leq 4\delta$. So by condition (2), with $f = g^{-1}h$, the isometry $g^{-1}h$ acts as the identity on D_p. Since an element of G is determined by its action on D_p, we have $h = g$.

Next assume that $g_0 := hgh^{-1} = g$. We claim that $d(g, h) = 0$, in which case $h = g$ follows from what has already been proved. To prove the claim, express $h(x) = Qx + q$ as in Lemma 1, where Q is orthogonal and $q \in V := axis(h)$. Taking the center of Q as the origin, let $W + a$ be the axis of g, where W is a linear subspace of $\mathbb{R}^n$. Using statement (1) of Lemma 3, $Q(W) + Qa + q = h(W + a) = h(axis(g)) = axis(g_0) = axis(g) = W + a$. This implies both (a) $Q(W) = W$ and (b) $(I - Q)a \in q + W$. But $V = axis(h) = k(axis(g)) = k(a + W)$, which, by the same reasoning as above, implies (c) $V = A(W)$, where A is the rotational part of k centered at the origin. We next prove, by contradiction, that $W = V$. Since subspaces V and W have the same dimension, assume that there exists a $w \in W \setminus V$. Let $w = v + v^{\perp}$, where $v \in V$ and $v^{\perp} \in V^{\perp}$ and let $x = w - Qw$. Then $x \in W$ because of statement (a), and $x \in V^{\perp}$ because $x = (v + v^{\perp}) - (Qv + Qv^{\perp}) = v^{\perp} - Qv^{\perp} \in V^{\perp}$. Hence, by statement (c), we know that A takes the element x of $V^{\perp}$ to an element of V. This contradicts $rot(k) < \sqrt{2}$. Now $W = V$ and, from statement (b), we have $(I - Q)a \in V$, which implies that $a \in V$ because $I - Q$ leaves both V and $V^{\perp}$ invariant and is non-singular when restricted to $V^{\perp}$. Hence $axis(g) = W + a = V + a = V = axis(h)$. ■

6. A CRYSTALLOGRAPHIC RESTRICTION AND THE PROOF OF BIEBERBACH'S THEOREM. Theorem 4 is an analog in $\mathbb{R}^n$ of the Crystallographic Restriction discussed in Section 4. In particular, if X is a crystal, then Theorems 2 and 4 eliminate the possibility of X possessing a k-fold rotation about a codimension two axis if $k \geq 13$. Also notice that Bieberbach's Theorem is an immediate corollary of Theorem 4 because the existence of translations in n linearly independent directions is guaranteed by Lemma 2.

Theorem 4. *If g is any non-identity element of a crystallographic group such that $rot(g) \leq 1/2$, then g must be a translation.*

Proof: By way of contradiction, assume that g is not a translation. Let δ and c be as in Lemma 4, and let $\varepsilon = \min(\delta, c/(4|trans(g)|))$. Consider the set C consisting of all conjugates $\bar{g} = kgk^{-1}$ of g in G such that

$$(1) \quad \bar{g} \neq g \qquad \text{and} \qquad (2) \quad rot(k) < \varepsilon.$$

The set C is not empty for the following reason. Since g is not a translation, $axis(g) \neq \mathbb{R}^n$. Lemma 2, with point p on $axis(g)$, then guarantees the existence of an isometry $k \in G$ such that $rot(k) < \varepsilon$ and the translational part of k does not lie in $axis(g)$. The latter condition implies that $axis(\bar{g})$ and $axis(g)$ are distinct because, by Lemma 3, $axis(\bar{g}) = k(axis(g))$. Since $axis(\bar{g}) \neq axis(g)$, also $\bar{g} \neq g$.

'Let

$$d = \inf_{\bar{g} \in C} d(g, \bar{g}) > c > 0,$$

the inequalities resulting directly from Lemma 4. The contradiction that will finish the proof is the existence of a $g_0 \in C$ such that $d(g, g_0) < d$. Let $h \in C$ be such that

$$d(g, h) \leq \tfrac{5}{4}d.$$

Then $g_0 = hgh^{-1}$ is such an element. It remains to show only that $g_0 \in C$ and that $d(g, g_0) < d$.

We first show that $g_0 \in C$. Because $h \in C$, we have $h \neq g$ and $h = kgk^{-1}$, where $rot(k) < \varepsilon \leq \delta$. Therefore $g_0 \neq g$ by Lemma 4. To verify the second condition in the definition of C, we show that there exists a $\bar{k} \in G$ such that $g_0 = \bar{k}g\bar{k}^{-1}$, where $rot(\bar{k}) < \varepsilon$. Since $g_0 = hgh^{-1} = (hg^{-1})g(hg^{-1})^{-1}$, statement (4) of Lemma 3 implies that $rot(hg^{-1}) \leq 2rot(g)rot(k) \leq rot(k) < \varepsilon$. Hence take $\bar{k} = hg^{-1}$.

To show that $d(g, g_0) < d$, let $V = axis(g)$ and $V' = axis(h)$. Let $p \in V$ and $p' \in V'$ be closest points on V and V', respectively. Further, let $\bar{V}$ denote the image of V under the translational part of h, and let $\bar{p} \in \bar{V}$ be a closest point to p on $\bar{V}$. If $trans(h) = 0$ then $\bar{p} = p$. Otherwise, since $h \in C$ express $h = kgk^{-1}$, and let α be the angle between $trans(g)$ and $trans(h)$. By elementary trigonometry we have $\sin(\alpha) \leq rot(k)$. Condition (2) in the definition of C and statement (2) of Lemma 3 yield

$$|p - \bar{p}| \leq |trans(h)|\sin(\alpha) \leq |trans(g)|rot(k) < \tfrac{1}{4}c \leq \tfrac{1}{4}d$$

$$|\bar{p} - p'| \leq |\bar{p} - p| + |p - p'| < \tfrac{1}{4}d + d(g,h) \leq \tfrac{3}{2}d.$$

If p'' is the image of $\bar{p}$ under the rotational part of h then, using statement (3) of Lemma 3,

$$|\bar{p} - p''| \leq |\bar{p} - p'|rot(h) < \tfrac{3}{2}d\, rot(g) \leq \tfrac{3}{4}d.$$

But $p \in axis(g)$ and p'', being in the image of $axis(g)$ under h, lies in $axis(g_0)$. Therefore

$$d(g, g_0) \leq |p - p''| \leq |p - \bar{p}| + |\bar{p} - p''| < \tfrac{1}{4}d + \tfrac{3}{4}d = d. \qquad \blacksquare$$

REFERENCES

1. L. Bieberbach, Über die Bewegungsgruppen der Euklidischen Raume, I, *Math. Ann.* 70 (1911), 297–336; II, Math. Ann. 72 (1912), 400–412.
2. H. Brown, R. Bülow, J. Neubüser, H. Wondratschek, and H. Zassenhaus, *Crystallographic Groups of Four-Dimensional Space*, Wiley, New York, 1978.
3. N. G. de Bruijn, Algebraic theory of Penrose's non-periodic tilings of the plane I, II, *Proc. Konink. Ned. Akad. Wetensch* A84 (1981), 39–66.
4. P. Buser, A geometric proof of Bieberbach's theorems on crystallographic groups, *L'Ensignment Mathématique* 31 (1985), 137–145.
5. L. S. Charlap, *Bieberbach Groups, and Flat Manifolds*, Springer-Verlag, New York, 1986.
6. B. Delone, N. Dolbilin, M. Shtogrin, and R. Galiulin, A local criterion for the regularity of a system of points, *Soviet Math. Dokl.* 17 (1976), 319–322.
7. E. S. Fedorov, Symmetry in the plane, *Zapiski Rus. Mineralog. Obščestva, Ser. 2* 28 (1891), 345–390.
8. C. Frobenius, Über die unzerlegbaren diskreten Bewegengsgruppen, *Sitzungsber. Akad. Wiss. Berlin* 29 (1911), 654–665.
9. A. Katz and M. Duneau, Quasiperiodic patterns and icosahedral symmetry, *J. Physique* 47 (1986), 181–196.
10. P. Kramer and R. Neri, On periodic and non-periodic space fillers of E^m obtained by projection, *Acta Cryst*. A40, 580–587.
11. S. Mozes, Tilings, substitution systems and dynamical systems generated by them, *J. Analyse Math.* 53 (1989), 139–186.
12. R. Penrose, The role of aesthetics in pure and applied mathematics, *Bull. Inst. Math. Applications* 10 (1974), 266–271.

13. C. Radin, Symmetry of tilings of the plane, *Bull. Amer. Math. Soc.* 29 (1993), 213–217.
14. A. Schoenflies, *Kristallsysteme und Kristallstruktur*, Teubner, Leipzig, 1981.
15. M. Senechal and J. Taylor, Quasicrystals: the view from Les Houches, *Math. Intelligencer* 12 (1990), 54–64.
16. D. Shechtman, I. Blech, D. Gratias, and J. Cahn, Metallic phase with long-range orientational order and no translational symmetry, *Phys. Rev. Lett.* 53 (1984), 1951–1954.
17. J. Wolf, *Spaces of Constant Curvature*, McGraw-Hill, New York, 1967.

Department of Mathematics
University of Florida
P.O. Box 118105, 358 Little Hall
Gainesville, FL 32611-8105
vince@math.ufl.edu

[redacted]

Integration Over Spheres and the Divergence Theorem for Balls

John A. Baker

The divergence theorem of Gauss asserts that if Ω is a nonempty bounded open subset of $\mathbf{R}^n (n \geq 2)$ with C^1 boundary $\partial\Omega$ and if $F\colon \overline{\Omega} \to \mathbf{R}^n$ is a C^1 function then

$$(\star) \qquad \int_{\overline{\Omega}} div F(x)\,dx = \int_{\partial\Omega} F(s) \cdot \nu(s)\,d\sigma(s)$$

where $\nu(s)$ is the "unit outward normal" to Ω at $s \in \partial\Omega$ and where integration on the right is with respect to the "$n-1$ dimensional surface measure" on $\partial\Omega$. A major obstacle in its formulation is to make sense of the right side of $(\star)$, i.e., to suitably define ν and integration over $\partial\Omega$.

The special case in which Ω is a ball (and $\partial\Omega$ a sphere) is of considerable interest because, for example, many fundamental properties of harmonic functions (including the mean value property, the weak maximum principle and the Poisson integral formula) follow from it in a relatively painless fashion; see, e.g., Chapter 1 of [1] and Chapter 2 of [4]. Even this case, when included in advanced calculus texts (usually with $n = 2$ or 3), is either dependent on a sophisticated and lengthy discussion of Stokes' theorem or on a definition of surface integral that involves improper integrals (in dimension $n - 1$), which are usually inadequately treated. Moreover the *rotational* (or *isometric*) invariance of integration over a sphere—in a sense (see the concluding remark) its characteristic property—is rarely mentioned in calculus texts.

The theory of harmonic functions in two real variables is addressed in many introductory analysis texts, presumably because it is virtually synonymous with analytic function theory (in one complex dimension). Such books are rarely concerned with harmonic functions of three or more variables, perhaps due to the lack of a suitably succinct, but sufficiently profound, version of the divergence theorem.

The aim of this article is to develop a utility-grade theory of integration over spheres and use it to formulate and prove the divergence theorem for balls in $\mathbf{R}^n$. Actually, we carry out this programme in detail only in case $\overline{\Omega} = B^n$—the closed ball in $\mathbf{R}^n$ centered at the origin with radius 1—and where we denote $\partial\Omega$ by S^{n-1} in deference to our topological friends. As we observe in §6, the case of an arbitrary ball can, without difficulty, be reduced to that of B^n.

Throughout this paper n denotes a fixed but arbitrary natural number. Unless otherwise indicated, $n \geq 2$. Here is an outline of our development.

(i) Define, in a simple way, an appropriate "integral" for continuous real valued functions on S^{n-1}, deduce salient properties thereof, including its rotational invariance and a "polar coordinates change of variable" formula (Theorem 1) and, with the aid of the gamma function, compute the integral of a polynomial over B^n and over S^{n-1}.

(ii) Use (i) to formulate and prove the divergence theorem for polynomial functions over B^n.
(iii) Use the higher-dimensional Weierstrass approximation theorem and (ii) to prove the divergence theorem for B^n.

1. BACKGROUND AND NOTATION. Let $\mathbf{Z}$ denote the integers. For $1 \le n \in \mathbf{Z}$ and $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n) \in \mathbf{R}^n$ let $x \cdot y = x_1 y_1 + \cdots + x_n y_n$ and $|x| = (x \cdot x)^{1/2}$. For $1 \le n \in \mathbf{Z}$ let $B^n = \{x \in \mathbf{R}^n : |x| \le 1\}$ and let $S^{n-1} = \{x \in \mathbf{R}^n : |x| = 1\}$, the boundary of B^n. If $A \subseteq \mathbf{R}^n$ let $A_0 = A \setminus \{0\}$. For $A \subseteq \mathbf{R}$ let $A_+ = \{x \in A : x \ge 0\}$. For $1 \le n \in \mathbf{Z}$ and $a, b \in \mathbf{R}$ with $0 \le a < b$, let $\mathbf{B}^n(a, b) = \{x \in \mathbf{R}^n : a \le |x| \le b\}$; if a real-valued function f is defined and Riemann integrable on $\mathbf{B}^n(a, b)$ we will sometimes denote the Riemann integral of f thereover by

$$\int_{a \le |x| \le b} f(x)\, dx.$$

For a given nonempty subset A of $\mathbf{R}^n (n \ge 1)$ and a bounded $f\colon A \to \mathbf{R}$ (or $\mathbf{R}^n$), let $\|f\|_A = \sup\{|f(x)| : x \in A\}$. We will need the following linear cases of the "change of variable theorem".

Proposition 1. *If* $0 \le a < b$, $f\colon \mathbf{B}^n(a, b) \to \mathbf{R}$ *is Riemann integrable on* $B^n(a, b)$, $0 < \rho \in \mathbf{R}$, and Q is a real orthogonal n × n matrix then

(i) $$\int_{a \le |x| \le b} f(x)\, dx = \rho^n \int_{a/\rho \le |y| \le b/\rho} f(\rho y)\, dy$$

and

(ii) $$\int_{a \le |x| \le b} f(x)\, dx = \int_{a \le |y| \le b} f(yQ)\, dy;$$

in particular,

(iii) $$\int_{a \le |x| \le b} f(x)\, dx = \int_{a \le |y| \le b} f(\epsilon_1 y_1, \dots, \epsilon_n y_n)\, d(y_1, \dots, y_n)$$

if $\epsilon_1, \dots, \epsilon_n \in \{1, -1\}$.

We will refer to property (ii) as *rotational* (or *isometric*) invariance. For most of what follows we will need this theorem only in case $f\colon \mathbf{R}^n \to \mathbf{R}$ and f is continuous on $\mathbf{R}^n_0$.

2. THE JOY OF INTEGRATION ON S^{n-1}. Suppose that $n \ge 2$ and $g\colon S^{n-1} \to \mathbf{R}$ is continuous. We aim to define an "integral" of g over S^{n-1}, which we denote by $\int_{S^{n-1}} g\, d\sigma_{n-1}$, in such a way that when $n = 2$ (respectively, $n = 3$) it conjures up arc length (respectively, surface area). Given such a g, define $\breve{g}\colon \mathbf{R}^n \to \mathbf{R}$ by

$$\breve{g}(x) = \begin{cases} g(|x|^{-1}x) & \text{for } x \in \mathbf{R}^n_0 \\ 0 & \text{for } x = 0. \end{cases}$$

Then $\breve{g}$ is continuous on $\mathbf{R}^n_0$, $\breve{g}(rs) = \breve{g}(s)$ for $s \in S^{n-1}$ and $0 < r \in \mathbf{R}$, and $\|g\|_{S^{n-1}} = \|\breve{g}\|_{R^n}$. It follows that $\breve{g}$ is Riemann integrable on B^n (or any other

subset of $\mathbf{R}^n$ having Jordan content). We may therefore legitimately define

$$\int_{S^{n-1}} g\, d\sigma_{n-1} := n\int_{B^n} \breve{g}(x)\, dx; \tag{1}$$

it will sometimes be convenient to use "Leibnizian" notation and write

$$\int_{S^{n-1}} g(s)\, d\sigma_{n-1}(s) \quad \text{or} \quad \int_{S^{n-1}} g(s_1,\dots,s_n)\, d\sigma_{n-1}(s_1,\dots,s_n)$$

instead of $\int_{S^{n-1}} g\, d\sigma_{n-1}$

One of the main payoffs of this definition is Theorem 1 below. Before discussing it we observe some simpler properties of our integral.

First note that if $V_n = \int_{B^n} 1\, dx$, the "n-dimensional volume" of B^n, and if $A_{n-1} = \int_{S^{n-1}} 1\, d\sigma_{n-1}$, the "$n - 1$ dimensional surface area" of S^{n-1}, then

$$A_{n-1} = nV_n \qquad \text{for } n \geq 2.$$

If g: $S^1 \to \mathbf{R}$ is continuous then

$$\int_{S^1} g d\sigma_1 = \int_0^{2\pi} g(\cos\theta, \sin\theta)\, d\theta$$

—the arc length integral—since (with the aid of polar coordinates)

$$\int_{B^2} g\left(\frac{x}{\sqrt{x^2+y^2}}, \frac{y}{\sqrt{x^2+y^2}}\right) d(x,y) = \int_0^{2\pi}\int_0^1 g(\cos\theta, \cos\theta) r\, dr\, d\theta.$$

It follows that $V_2 = \pi$ and $A_1 = 2\pi$, as expected. Here are some other elementary properties of integration over S^{n-1} that follow easily from the definition and Proposition 1.

Proposition 2. *If g, h: $S^{n-1} \to \mathbf{R}$ are continuous, α, $\beta \in \mathbf{R}$, and Q is a real orthogonal $n \times n$ matrix, then*

(i) $\displaystyle\int_{S^{n-1}} (\alpha g + \beta h)\, d\sigma_{n-1} = \alpha\int_{S^{n-1}} g\, d\sigma_{n-1} + \beta\int_{S^{n-1}} h\, d\sigma_{n-1}$,

(ii) $\displaystyle\int_{S^{n-1}} g\, d\sigma_{n-1} > 0$ *if $g(s) \geq 0$ for all $s \in S^{n-1}$ and $g(s_0) > 0$ for some $s_0 \in S^{n-1}$*,

(iii) $\displaystyle\left|\int_{S^{n-1}} g\, d\sigma_{n-1}\right| \leq \int_{S^{n-1}} |g(s)|\, d\sigma_{n-1}(s) \leq \|g\|_{S^{n-1}} A_{n-1}$, *and*

(iv) $\displaystyle\int_{S^{n-1}} g(s)\, d\sigma_{n-1}(s) = \int_{S^{n-1}} g(sQ)\, d\sigma_{n-1}(s)$—*"rotational invariance"*.

The following assertion is a useful generalization of the 2-dimensional "polar coordinate change of variable theorem".

Theorem 1. *Suppose $0 \leq a < b$ and f: $B^n(a,b) \to \mathbf{R}$ is continuous. Then*

$$\int_{a \leq |x| \leq b} f(x)\, dx = \int_a^b r^{n-1}\left(\int_{S^{n-1}} f(rs)\, d\sigma_{n-1}(s)\right) dr. \tag{2}$$

Moreover, if

$$\varphi(r) = \int_{a \le |x| \le r} f(x)\,dx \quad \text{for } a \le r \le b,$$

then φ is continuously differentiable and

$$\varphi'(r) = r^{n-1} \int_{S^{n-1}} f(rs)\,d\sigma_{n-1}(s) \quad \text{for } a \le r \le b,$$

Proof: Assume for the moment that $a > 0$ and let $\epsilon > 0$. Choose $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ whenever $x, y \in B^n(a, b)$ and $|x - y| \le \delta$. Suppose that $a \le r < r + h \le b$ and $h \le \delta$. Then

$$\left| \int_{r \le |x| \le r+h} f(x)\,dx - \int_{r \le |x| \le r+h} f(r|x|^{-1}x)\,dx \right|$$
$$= \left| \int_{r \le |x| \le r+h} \{ f(x) - f(r|x|^{-1}x) \}\,dx \right| \le \epsilon\big((r+h)^n - r^n\big)V_n$$

since $|x - r|x|^{-1}x| = |x| - r \le h \le \delta$ if $r \le |x| \le r + h$. But

$$\int_{r \le |x| \le r+h} f(r|x|^{-1}x)\,dx = \int_{0 < |x| \le r+h} f(r|x|^{-1}x)\,dx - \int_{0 < |x| \le r} f(r|x|^{-1}x)\,dx$$
$$= (r+h)^n \int_{0 < |y| \le 1} f(r|y|^{-1}y)\,dy - r^n \int_{0 < |y| \le 1} f(r|y|^{-1}y)\,dy$$
$$= \frac{(r+h) - r^n}{n} \int_{S^{n-1}} f(rs)\,d\sigma_{n-1}(s).$$

By the last two observations and the definition of φ, for $a \le r < r + h \le b$,

$$\left| \frac{\varphi(r+h) - \varphi(r)}{h} - \left[\frac{(r+h)^n - r^n}{nh} \right] \int_{S^{n-1}} f(rs)\,d\sigma_{n-1}(s) \right| \le \epsilon \left[\frac{(r+h)^n - r^n}{h} \right] V_n$$

and hence

$$\left| \frac{\varphi(r+h) - \varphi(r)}{h} - r^{n-1} \int_{S^{n-1}} f(rs)\,d\sigma_{n-1}(s) \right|$$
$$\le \epsilon \left[\frac{(r+h)^n - r^n}{h} \right] V_n + \left[\frac{(r+h)^n - r^n}{nh} - r^{n-1} \right] \left| \int_{S^{n-1}} f(rs)\,d\sigma_{n-1}(s) \right|.$$

It follows that φ is differentiable on $[a, b]$ and

$$\varphi'(r) = r^{n-1} \int_{S^{n-1}} f(rs)\,d\sigma_{n-1}(s) \quad \text{for } a \le r \le b.$$

The uniform continuity of f implies that φ' is continuous. By the Fundamental Theorem of Calculus,

$$\int_{a \le |x| \le b} f(x)\,dx = \varphi(b) - \varphi(a) = \int_a^b \varphi'(r)\,dr = \int_a^b r^{n-1} \int_{S^{n-1}} f(rs)\,d\sigma_{n-1}(s)\,dr.$$

Let a tend to zero to complete the proof. ■

Corollary 1. *Suppose that $f\colon \mathbf{R}^n \to \mathbf{R}$, f is continuous, and f is homogeneous of degree $\rho > 0$, i.e., $f(rx) = r^\rho f(x)$ whenever $x \in \mathbf{R}^n$ and $r \ge 0$. Then*

$$\int_{B^n} f(x)\,dx = \frac{1}{\rho + n} \int_{S^{n-1}} f(s)\,d\sigma_{n-1}(s).$$

Proof: By Theorem 1,

$$\int_{B^n} f(x)\,dx = \int_0^1 r^{n-1}\left(\int_{S^{n-1}} f(rs)\,d\sigma_{n-1}(s)\right)dr$$
$$= \left(\int_0^1 r^{p+n-1}\,dr\right)\int_{S^{n-1}} f(s)\,d\sigma_{n-1}(s) = \frac{1}{p+n}\int_{S^{n-1}} f(s)\,d\sigma_{n-1}(s). \quad \blacksquare$$

Remarks. The author suspects that (1) is folkloric, and likely has appeared as a theorem rather than a definition. In exercise 6 on page 175 of [7], Rudin outlines measure-theoretic generalizations of these ideas.

If for a continuous $g\colon S^{n-1} \to \mathbf{R}$ we define

$$g^{\#}(x) = \begin{cases} |x| g(|x|^{-1}x) & \text{if } 0 \neq x \in \mathbf{R}^n \\ 0 & \text{if } x = 0 \in \mathbf{R}^n \end{cases}$$

then $g^{\#}$ is continuous on $\mathbf{R}^n$ and homogeneous of degree 1; hence, by the corollary,

$$\int_{S^{n-1}} g(s)\,d\sigma_{n-1}(s) = (n+1)\int_{B^n} g^{\#}(x)\,dx.$$

This could have served as a definition of

$$\int_{S^{n-1}} g\,d\sigma_{n-1},$$

thereby allowing the definition of integration over S^{n-1} to depend only on the Riemann integrability of a *continuous* function on B^n (our $g^{\#}$ is continuous on $\mathbf{R}^n$ but $\check{g}$ is discontinuous at 0 unless $g \equiv 0$).

Suppose $g\colon S^{n-1} \to \mathbf{R}$ is continuous and f is a continuous extension of g to $\mathbf{R}^n$. For $0 \le r_1 < r_2$ it is natural to call

$$\frac{1}{V_n(r_2^n - r_1^n)}\int_{r_1 \le |x| \le r_2} f(x)\,dx$$

the *average value* of f on $B^n(r_1, r_2)$; denote it by $av(f, r_1, r_2)$. It is also natural to call

$$\frac{1}{A_{n-1}}\int_{S^{n-1}} g\,d\sigma_{n-1}$$

the *average* of g on S^{n-1}. It follows from Theorem 1 that

$$A_{n-1}^{-1}\int_{S^{n-1}} g\,d\sigma_{n-1} = \lim_{\substack{r_1 < r_2 \\ r_2 - r_1 \to 0}} av(f, r_1, r_2).$$

In fact the plausibility of this and of Theorem 1 supplied the intuition for our definition.

3. AN INTEGRATION FORMULA INVOLVING THE GAMMA FUNCTION.

It will be crucial to integrate polynomials over B^n and over S^{n-1}.

Given $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbf{Z}_+^n$ (such an α is called a *multi-index* and $|\alpha| := \sum_{j=1}^n \alpha_j$ is its *order*), define

$$p_\alpha(x) = x_1^{\alpha_1} \cdots x_n^{\alpha_n} \quad \text{for } x = (x_1, \ldots, x_n) \in \mathbf{R}^n$$

(with the convention $0^0 = 1$). Such a p_α will be called a *monomial.* Every polynomial (function) from $\mathbf{R}^n$ to $\mathbf{R}$ is a finite linear combination of monomials.

Suppose $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbf{Z}_+^n$. Since p_α is continuous and homogeneous of degree $|\alpha|$, Corollary 1 ensures that,

$$\int_{S^{n-1}} p_\alpha(s)\, d\sigma_{n-1}(s) = (|\alpha| + n) \int_{B^n} p_\alpha(x)\, dx. \tag{3}$$

Moreover, by (ii) of Proposition 1, if $\beta = (\beta_1, \dots, \beta_n)$ is a permutation of $(\alpha_1, \dots, \alpha_n)$ then

$$\int_{B^n} p_\alpha(x)\, dx = \int_{B^n} p_\beta(x)\, dx,$$

and hence

$$\int_{S^{n-1}} p_\alpha\, d\sigma_{n-1} = \int_{S^{n-1}} p_\beta\, d\sigma_{n-1}.$$

If one of $\alpha_1, \dots, \alpha_n$ is odd we claim that

$$\int_{S^n} p_\alpha(x)\, dx = 0 \quad \text{and} \quad \int_{S^{n-1}} p_\alpha(s)\, d\sigma_{n-1}(s) = 0.$$

For example, if α_1 is odd then by (iii) of Proposition 1,

$$\int_{B^n} p_\alpha(x)\, dx = \int_{B^n} (-x_1)^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}\, d(x_1, \dots, x_n) = -\int_{B^n} p_\alpha(x)\, dx,$$

so that

$$\int_{B^n} p_\alpha(x)\, dx = 0$$

and therefore, by (3),

$$\int_{B^n} p_\alpha(s)\, d\sigma_{n-1}(s) = 0.$$

The task of integrating a polynomial over B^n (or S^{n-1}) is therefore reduced to that of finding

$$\int_{B^n} p_\alpha(x)\, dx$$

in case $\alpha = (2\beta_1, \dots, 2\beta_n)$ with $\beta_1, \dots, \beta_n \in \mathbf{Z}_+^n$.

As it turns out, we can do a bit better with the help of Euler's gamma function. We will use the well-known fact from advanced calculus (see eg. **[3]**, page 294 or **[5]**, page 484) that

$$\int_0^1 t^\lambda (1-t)^\mu\, dt = \frac{\Gamma(\lambda + 1)\Gamma(\mu + 1)}{\Gamma(\lambda + \mu + 2)} \quad \text{for } -1 < \lambda, \mu \in \mathbf{R} \tag{4}$$

where Γ is Euler's gamma function. Recall that

$$\Gamma(x+1) = x\Gamma(x) > 0 \qquad \text{for } 0 < x \in \mathbf{R}, \text{ and} \tag{5}$$

$$\Gamma(k) = (k-1)! \qquad \text{for } 0 < k \in \mathbf{Z}. \tag{6}$$

For $0 \le \gamma_1, \dots, \gamma_n \in \mathbf{R}_+$, define

$$I_n(\gamma_1, \dots, \gamma_n) = \int_{B^n} (x_1^2)^{\gamma_1} \cdots (x_n^2)^{\gamma_n}\, d(x_1, \dots, x_n);$$

by Corollary 1

$$I_n(\gamma_1, \dots, \gamma_n) = (2\gamma_1 + \cdots + 2\gamma_n + n)^{-1} \int_{S^{n-1}} (s_1^2)^{\gamma_1} \cdots (s_n^2)^{\gamma_n} \, d\sigma_{n-1}(s_1, \dots, s_n).$$

If λ, $\mu \in \mathbf{R}_+$ then

$$\begin{aligned} I_2(\lambda, \mu) &= \int_{B^2} (x^2)^{\lambda} (y^2)^{\mu} \, d(x, y) = 4\int_0^1 \left(\int_0^{\sqrt{1-x^2}} x^{2\lambda} y^{2\mu} \, dy \right) dx \\ &= 4\int_0^1 x^{2\lambda} \frac{(1 - x^2)^{\frac{2\mu+1}{2}}}{2\mu + 1} \, dx = \frac{4}{2\mu + 1} \int_0^1 t^{\lambda} (1 - t)^{\mu + 1/2} \frac{dt}{2\sqrt{t}} \\ &= \frac{2}{2\mu + 1} \int_0^1 t^{\lambda - 1/2} (1 - t)^{\mu + 1/2} \, dt \\ &= \frac{1}{\mu + 1/2} \frac{\Gamma(\lambda + 1/2)\Gamma(\mu + 3/2)}{\Gamma(\lambda + \mu + 2)} \qquad \text{(by (4))} \\ &= \frac{\Gamma(\lambda + 1/2)\Gamma(\mu + 1/2)}{\Gamma(\lambda + \mu + 2)} \qquad \text{(by (5))}. \end{aligned}$$

Thus we have

$$I_2(\gamma_1, \gamma_2) = \frac{\Gamma(\gamma_1 + 1/2)\Gamma(\gamma_2 + 1/2)}{\Gamma(\gamma_1 + \gamma_2 + 2)} \quad \text{for } \gamma_1, \gamma_2 \in \mathbf{R}_+. \tag{7}$$

Now suppose that $n \geq 3$ and $\gamma_1, \dots, \gamma_n \in \mathbf{R}_+$. By integrating first with respect to x_n, using Theorem 1 (with n replaced by $n - 1$), and changing variables we surmize that

$$\begin{aligned} &I_n(\gamma_1, \dots, \gamma_n) \\ &= 2\int_{B^{n-1}} (x_1^2)^{\gamma_1} \cdots (x_{n-1}^2)^{\gamma_{n-1}} \frac{[1 - (x_1^2 + \cdots x_{n-1}^2)]^{\frac{2\gamma_n+1}{2}}}{2\gamma_n + 1} \, d(x_1, \dots, x_{n-1}) \\ &= \frac{2}{2\gamma_n + 1} \int_0^1 r^{n-2} \left[\int_{S^{n-2}} (r^2)^{\gamma_1 + \cdots + \gamma_{n-1}} (s_1^2)^{\gamma_1} \cdots (s_{n-1}^2)^{\gamma_{n-1}} \right. \\ &\qquad\qquad \left. \times (1 - r^2)^{\gamma_n + 1/2} \, d\sigma_{n-2}(s) \right] dr \\ &= \frac{2}{2\gamma_n + 1} \left[\int_0^1 (r^2)^{\gamma_1 + \cdots + \gamma_{n-1} - 1 + n/2} (1 - r^2)^{\gamma_n + 1/2} \, dr \right] \\ &\qquad \times \int_{S^{n-2}} (s_1^2)^{\gamma_1} \cdots (s_{n-1}^2)^{\gamma_{n-1}} \, d\sigma_{n-2}(s) \\ &= \frac{2}{2\gamma_n + 1} \left[\int_0^1 t^{\gamma_1 + \cdots + \gamma_{n-1} - 1 + n/2} (1 - t)^{\gamma_n + 1/2} \frac{dt}{2\sqrt{t}} \right] \\ &\qquad \times \int_{S^{n-2}} (s_1^2)^{\gamma_1} \cdots (s_{n-1}^2)^{\gamma_{n-1}} \, d\sigma_{n-2}(s). \end{aligned}$$

But, according to the Corollary 1,

$$\int_{S^{n-2}} (s_1^2)^{\gamma_1} \cdots (s_{n-1}^2)^{\gamma_{n-1}} \, d\sigma_{n-2}(s)$$
$$= (2\gamma_1 + \cdots + 2\gamma_{n-1} + n - 1) I_{n-1}(\gamma_1, \ldots, \gamma_{n-1})$$

and, by (4),

$$\int_0^1 t^{\gamma_1 + \cdots + \gamma_{n-1} - 3/2 + n/2} (1-t)^{\gamma_n + 1/2} \, dt$$
$$= \frac{\Gamma(\gamma_1 + \cdots + \gamma_{n-1} - 1/2 + n/2)\Gamma(\gamma_n + 3/2)}{\Gamma(\gamma_1 + \cdots + \gamma_n + 1 + n/2)}.$$

Thus,

$$I_n(\gamma_1, \ldots, \gamma_n)$$
$$= \frac{1}{(2\gamma_n + 1)} \frac{\Gamma(\gamma_1 + \cdots + \gamma_{n-1} - 1/2 + n/2)\Gamma(\gamma_n + 3/2)}{\Gamma(\gamma_1 + \cdots + \gamma_n + 1 + n/2)}$$
$$\cdot (2\gamma_1 + \cdots + 2\gamma_{n-1} + n - 1) I_{n-1}(\gamma_1, \ldots, \gamma_{n-1})$$
$$= \frac{(\gamma_1 + \cdots + \gamma_{n-1} - 1/2 + n/2)\Gamma(\gamma_1 + \cdots + \gamma_{n-1} - 1/2 + n/2)}{\Gamma(\gamma_1 + \cdots + \gamma_n + 1 + n/2)}$$
$$\cdot \frac{\Gamma(\gamma_n + 3/2)}{(\gamma_n + 1/2)} \cdot I_{n-1}(\gamma_1, \ldots, \gamma_{n-1})$$
$$= \frac{\Gamma(\gamma_1 + \cdots + \gamma_{n-1} + 1/2 + n/2)\Gamma(\gamma_n + 1/2)}{\Gamma(\gamma_1 + \cdots + \gamma_n + 1 + n/2)} I_{n-1}(\gamma_1, \ldots, \gamma_{n-1})$$

where the punch-line needed (5).

It follows by induction, with the aid of (4) and (7), that

$$I_n(\gamma_1, \ldots, \gamma_n) = \frac{\Gamma(\gamma_1 + 1/2) \cdots \Gamma(\gamma_n + 1/2)}{\Gamma(\gamma_1 + \cdots + \gamma_n + 1 + n/2)} \quad \text{for } n \geq 2 \text{ and } \gamma_1, \ldots, \gamma_n \in \mathbf{R}_+. \tag{8}$$

4. THE DIVERGENCE THEOREM FOR B^n; A PROOF FOR POLYNOMIALS. A function f: $\mathbf{B}^n \to \mathbf{R}$ is said to be C^1 (continuously differentiable) on B^n provided there is a $\Delta > 0$ and a continuously differentiable function $\chi : \{ x \in \mathbf{R}^n : |x| < 1 + \Delta\} \to \mathbf{R}$ such that $f(x) = \chi(x)$ for all $x \in B^n$; in this case we define $\partial_i f(x) := \partial_i \chi(x)$ for $x \in B^n$ and $1 \leq i \leq n$ (where $\partial_i = \partial/\partial x_i$). Given such Δ and χ, choose a continuously differentiable function φ: $[0, +\infty) \to \mathbf{R}$ such that

$$\varphi(t) = 1 \text{ when } 0 \leq t \leq 1 + \Delta/3 \quad \text{and} \quad \varphi(t) = 0 \text{ when } 1 + \frac{2\Delta}{3} \leq t \in \mathbf{R}.$$

Then one may (without embarrasment) define f_0: $\mathbf{R}^n \to \mathbf{R}$ by decreeing that

$$f_0(x) = \begin{cases} \varphi(|x|)\chi(x) & \text{if } |x| < 1 + \Delta \\ 0 & \text{if } |x| > 1 + (2/3)\Delta \end{cases}$$

and conclude that f_0 is C^1 on $\mathbf{R}^n$ and $f_0(x) = f(x)$ for $x \in B^n$. In summary, a C^1 function from B^n to $\mathbf{R}$ is the restriction to B^n of a C^1 function from $\mathbf{R}^n$ to $\mathbf{R}$ (with compact support if desired).

Suppose that F: $B^n \to \mathbf{R}^n$, say

$$F(x) = (f_1(x), \dots, f_n(x)) \qquad \text{for } x \in B^n.$$

We say that f is C^1 on B^n provided every f_j is C^1 on B^n. In this case define

$$div F(x) = \sum_{i=1}^{n} \partial_i f_i(x) \qquad \text{for } x \in B^n;$$

the function $div F$ is called the *divergence* of F.

The divergence theorem for B^n asserts that if F: $B^n \to \mathbf{R}^n$ is C^1 then

$$\int_{B^n} div F(x)\, dx = \int_{S^{n-1}} F(s) \cdot s\, d\sigma_{n-1}(s).$$

With the notation of the preceding paragraph, this is equivalent to asserting that

$$\int_{B^n} \partial_i f(x)\, dx = \int_{S^{n-1}} f(s_1, \dots, s_n) s_i\, d\sigma_{n-1}(s_1, \dots, s_n), \quad 1 \le i \le n.$$

It therefore suffices to prove that, for any C^1 function f: $B^n \to \mathbf{R}$,

$$\int_{B^n} \partial_i f(x)\, dx = \int_{S^{n-1}} f(s_1, \dots, s_n) s_i\, d\sigma_{n-1}(s_1, \dots, s_n), \quad 1 \le i \le n.$$

The rotation invariance of integration (over B^n and over S^{n-1}) further implies that one i is as good as another, that is, the divergence theorem for B^n is equivalent to

Proposition 3. *For every C^1 function f: $\mathbf{R}^n \to \mathbf{R}$,*

$$\int_{B^n} \partial_n f(x)\, dx = \int_{S^{n-1}} f(s_1, \dots, s_n) s_n\, d\sigma_{n-1}(s_1, \dots, s_n). \tag{9}$$

Our strategy of proof is to apply the Weierstrass approximation theorem to reduce consideration to the case in which f is a polynomial function. The polynomial case easily reduces to the monomial case by linearity (of ∂_n, the dot product, and integration over both B^n and S^{n-1}).

Suppose that $(\alpha_1, \dots, \alpha_n) \in \mathbf{Z}_+^n$, and $f(x) = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ for all $x = (x_1, \dots, x_n) \in \mathbf{R}^n$.

Case 1. $\alpha_n = 0$: In this case, $\partial_n f \equiv 0$ and, by the "oddness" observation of §3,

$$\int_{S^{n-1}} f(s_1, \dots, s_n) s_n\, d\sigma_{n-1}(s_1, \dots, s_n) = \int_{S^{n-1}} s_1^{\alpha_1} \cdots s_{n-1}^{\alpha_{n-1}} s_n\, d\sigma_{n-1}(s_1, \dots, s_n) = 0.$$

Case 2. *One of* $\alpha_1, \dots, \alpha_{n-1}$ *is odd*, or, $\alpha_n > 0$ *and* α_n *is even*: In this case,

$$\int_{B^n} \partial_n f(x)\, dx = \alpha_n \int_{B^n} x_1^{\alpha_1} \cdots x_{n-1}^{\alpha_{n-1}} x_n^{\alpha_n - 1}\, d(x_1, \dots, x_n) = 0$$

because one of the exponents is odd (see §3), and, for the same reason,

$$\int_{S^{n-1}} f(s_1, \dots, s_n) s_n\, d\sigma_{n-1}(s_1, \dots, s_n)$$

$$= \int_{S^{n-1}} s_1^{\alpha_1} \cdots s_{n-1}^{\alpha_{n-1}} s_n^{\alpha_n + 1}\, d\sigma_n(s_1, \dots, s_n) = 0.$$

Case 3. *$\alpha_1, \dots, \alpha_{n-1}$ are even and α_n is odd*: Choose $\beta_1, \dots, \beta_n \in \mathbf{Z}_+$ such that $\alpha_k = 2\beta_k$ for $1 \le k \le n-1$ and $\alpha_n = 2\beta_n + 1$. Then

$$\int_{B^n} \partial_n f(x)\,dx = (2\beta_n + 1)\int_{B^n} x_1^{2\beta_1} \cdots x_n^{2\beta_n}\,d(x_1, \dots, x_n)$$
$$= (2\beta_n + 1)\frac{\Gamma(\beta_1 + 1/2) \cdots \Gamma(\beta_n + 1/2)}{\Gamma(\beta_1 + \cdots + \beta_n + 1 + n/2)}$$

according to (3) and (8). By (3), (8), and (5),

$$\int_{S^{n-1}} f(s_1, \dots, s_n) s_n\,d\sigma_{n-1}(s_1, \dots, s_n)$$
$$= \int_{S^{n-1}} s_1^{2\beta_1} \cdots s_{n-1}^{2\beta_{n-1}} s_n^{2\beta_n + 2}\,d\sigma_{n-1}(s_1, \dots, s_n)$$
$$= (2\beta_1 + \cdots + 2\beta_{n-1} + 2\beta_n + 2 + n) I_n(\beta_1, \dots, \beta_n + 1)$$
$$= 2(\beta_1 + \cdots + \beta_n + 1 + n/2)$$
$$\times \frac{\Gamma(\beta_1 + 1/2) \cdots \Gamma(\beta_{n-1} + 1/2)\Gamma(\beta_n + 3/2)}{\Gamma(\beta_1 + \cdots + \beta_n + 2 + n/2)}$$
$$= 2\frac{\Gamma(\beta_1 + 1/2) \cdots \Gamma(\beta_{n-1} + 1/2)(\beta_n + 1/2)\Gamma(\beta_n + 1/2)}{\Gamma(\beta_1 + \cdots + \beta_n + 1 + n/2)}$$
$$= \int_{B^n} \partial_n f(x)\,dx.$$

We have verified Proposition 3 in case f is a polynomial function.

5. COMPLETING THE PROOF WITH THE HELP OF WEIERSTRASS. We will use the following variant of the

Weierstrass Approximation Theorem. *If K is a nonempty compact subset of $\mathbf{R}^n$, $n \ge 1$, f: $K \to \mathbf{R}$, f is continuous, and $\epsilon > 0$, then there exists a polynomial p: $\mathbf{R}^n \to \mathbf{R}$ such that*

$$|f(x) - p(x)| < \epsilon \qquad \textit{for all } x \in \mathbf{K}.$$

This theorem can be proved in many ways. The Stone-Weierstrass Theorem (see, e.g., **[8]**, page 210) yields it almost immediately and the celebrated method of Bernstein can be adapted to higher dimensions (see, e.g., **[2]**, page 122). In fact Weierstrass' proof (see Chapter 59 of **[6]**) can also be modified, in a natural way, to apply when $n > 1$.

C6mpletion of the Proof of Proposition 3. Suppose that f: $B^n \to \mathbf{R}$, f is C^1 and $\epsilon > 0$. Assume, without loss of generality, that f is in fact defined and C^1 on $\mathbf{R}^n$.

For $x = (x_1, \dots, x_n) \in \mathbf{R}^n$,

$$f(x) = f(0) + (f(x_1, \dots, x_{n-1}, 0) - f(0)) \tag{10}$$
$$+ (f(x_1, \dots, x_n) - f(x_1, \dots, x_{n-1}, 0))$$
$$= f(0) + (f(x_1, \dots, x_{n-1}, 0) - f(0)) + \int_0^{x_n} \partial_n f(x_1, \dots, x_{n-1}, t)\,dt.$$

Choose polynomials $p_0\colon \mathbf{R}^{n-1} \to \mathbf{R}$ and $p_1\colon \mathbf{R}^n \to \mathbf{R}$ such that

$$|f(x_1,\dots,x_{n-1},0) - f(0) - p_0(x_1,\dots,x_{n-1})| < \epsilon/7 \tag{11}$$

and

$$|\partial_n f(x_1,\dots,x_n) - p_1(x_1,\dots,x_n)| < \epsilon/7 \quad \text{whenever } (x_1,\dots,x_n) \in B^n. \tag{12}$$

Now, for $x_1,\dots,x_n \in \mathbf{R}$, define

$$p(x_1,\dots,x_n) = f(0) + p_0(x_1,\dots,x_{n-1}) + \int_0^{x_n} p_1(x_1,\dots,x_{n-1},t)\,dt \tag{13}$$

and note that p is a polynomial. By (10)–(13), for all $x = (x_1,\dots,x_n) \in B^n$,

$$\begin{aligned}|f(x) - p(x)| &\le \epsilon/7 + \left|\int_0^{x_n} \partial_n f(x_1,\dots,x_{n-1},t) - p_1(x_1,\dots,x_{n-1},t)\,dt\right|\\ &\le \epsilon/7 + |x_n|\epsilon/7\end{aligned}$$

so that

$$|f(x) - p(x)| < \epsilon. \tag{14}$$

By (13), $\partial_n p = p_1$; hence, by (12),

$$|\partial_n f(x) - \partial_n p(x)| < \epsilon/7 \qquad \text{for all } x \in B^n. \tag{15}$$

From §4,

$$\int_{B^n} \partial_n p(x)\,dx = \int_{S^{n-1}} p(s_1,\dots,s_n)s_n\,d\sigma_{n-1}(s_1,\dots,s_n).$$

Thus, by (14) and (15),

$$\begin{aligned}&\left|\int_{B^n} \partial_n f(x)\,dx - \int_{S^{n-1}} f(s_1,\dots,s_n)s_n\,d\sigma_{n-1}(s_1,\dots,s_n)\right|\\ &\le \left|\int_{B^n} \partial_n f(x)\,dx - \int_{B^n} \partial_n p(x)\,dx\right|\\ &\quad + \left|\int_{S^{n-1}} p(s_1,\dots,s_n)s_n\,d\sigma_{n-1}(s_1,\dots,s_n) - \int_{S^{n-1}} f(s_1,\dots,s_n)s_n\,d\sigma_{n-1}(s_1,\dots,s_n)\right|\\ &\le (\epsilon/7)V_n + \epsilon A_{n-1}.\end{aligned}$$

Since ϵ was chosen arbitrarily, the proof of Proposition 3 is complete. ■

The divergence theorem for B^n has therefore been established.

6. FURTHER REMARKS.

1. The divergence theorem can be extended to arbitrary balls in $\mathbf{R}^n$ by appropriately defining integration over arbitrary spheres. This can be done as follows.
   Given $p \in \mathbf{R}^n$ and $\rho > 0$, let $B(p,\rho) = \{x \in \mathbf{R}^n\colon |x-p| \le \rho\}$ and $S(p,\rho) = \{x \in \mathbf{R}^n\colon |x-p| = \rho\}$. For a continuous $g\colon S(p,\rho) \to \mathbf{R}$ define

$$\begin{aligned}\int_{S(p,\rho)} g\,d\sigma_{n-1} &:= \rho^{n-1}\int_{S^{n-1}} g(p+\rho s)\,d\sigma_{n-1}(s)\\ &= n\rho^{n-1}\int_{B_0^n} g\left(p + \rho|x|^{-1}x\right)dx.\end{aligned}$$

2. By (8),

$$V_n = \frac{\Gamma(1/2)^n}{\Gamma(1+n/2)} = \frac{2}{n}\frac{\Gamma(1/2)^n}{\Gamma(n/2)} \quad \text{for } n \geq 2.$$

In particular, $\pi = V_2 = \Gamma(1/2)^2$; i.e.,

$$\Gamma(1/2) = \sqrt{\pi}.$$

Hence

$$V_n = \frac{2}{n}\frac{\pi^{n/2}}{\Gamma(n/2)} \quad \text{for } n \geq 2.$$

3. Riemann integration and Jordan content can be defined in the S^{n-1} setting as follows.
Given $A \subseteq S^{n-1}$, call A *contented* provided $\check{A} := \{rs : s \in A, 0 \leq r \leq 1\}$ has Jordan content and in this case define $\sigma_{n-1}(A) = n\mu_n(\check{A})$, where μ_n denotes Jordan content in $\mathbf{R}^n$. If $g: S^{n-1} \to \mathbf{R}$, let us say that g is (Riemann) *integrable* on A provided $\check{g}$ is Riemann integrable on $\check{A}$, in which case decree that

$$\int_A g\, d\sigma_{n-1} := n\int_{\check{A}} g(|x|^{-1}x)\, dx.$$

4. More ambitiously, one can define

$$\int_{S^{n-1}} g\, d\sigma_{n-1} := n\int_{B^n} g(|x|^{-1}x)\, dx$$

for any Borel measurable $g: S^{n-1} \to \mathbf{R}_+$ (this is the gist of exercise 6, Chapter 8 of **[7]**) and thereby construct a rotation-invariant regular Borel measure on S^{n-1}. This can also be accomplished by applying the Riesz Representation Theorem (see **[8]**, page 352) to extend our humble integral. In fact, a theorem of Banach (see **[9]**, pp. 314–319, or **[8]**, pp. 361–370) implies that there is a unique regular Borel probability measure on S^{n-1} that is rotation invariant.

REFERENCES

1. S. Axler, P. Bourdon, and W. Ramey, *Harmonic Function Theory*, Springer-Verlag, New York, 1992.
2. P. J. Davis, *Interpolation and Approximation*, Dover, New York, 1975.
3. A. Friedman, *Advanced Calculus*, Holt, Rinehart and Winston, New York, 1971.
4. D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.
5. W. Kaplan, *Advanced Calculus*, 4th Ed., Addison-Wesley, Reading, Mass., 1991.
6. T. W. Körner, *Fourier Analysis*, Cambridge University Press, Cambridge, 1989.
7. W. Rudin, *Real and Complex Analysis*, 3rd Ed., McGraw-Hill, New York, 1987.
8. H. L. Royden, *Real Analysis*, 3rd Ed., Macmillan, New York, 1988.
9. S. Saks, *Theory of the Integral*, 2nd revised Ed., Dover, New York, 1964

Department of Pure Mathematics
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada
jabaker@math.uwaterloo.ca

Math Lingo vs. Plain English: Double Entendre

Reuben Hersh

Once upon a time, when I was a teaching assistant, teaching a class of the kind mockingly called "Math for Poets," an obnoxious freshman said to me, "Zero isn't a number."

I have forgotten my answer, but I remember finding her remark a shocking expression of profound ignorance.

Years later, it dawned on me—she was right!

If I say "I own a number of calculus books" or "I have a number of friends at the Courant Institute," I don't mean *zero* books or *zero* friends. I don't even mean *one* book or *one* friend. I mean two or more. *That's* what "number" means in plain English. I read recently that the famous phenomenologist Edmund Husserl meant by "number" something greater or equal to 2. So did Plato.

In mathematical talk, "number" has several meanings. None is the plain English meaning. The ordinary math teacher, like me back then, is so deeply embedded in math lingo that he/she doesn't notice the inconsistency. But the inconsistency can confuse students.

I say "math lingo," not language. It's a jargon, a semidialect of English (or some other natural language), not a complete language. You can't say "I have a headache" or "You bore me" in math lingo.

In math lingo, a straight line is the simplest example of a curve. In plain English, quite otherwise: a straight line isn't a curve, and a curve isn't a straight line.

In English, what we call a "line segment" is just a "line." What we call a "line" is "an infinite line." "Difference," "product," "factor," "prime" all have different meanings in plain English and in math lingo. I may ask a student, "If you subtract zero from zero, what's the difference?" While answering math-linguistically, "zero," she may be thinking, plain-Englishly, "That's right! Who cares? What's the difference?"

In English, "adding" increases what you've got. In math lingo, it may increase it, decrease it or neither, depending on whether you happen to be adding something positive, negative or zero.

Correspondingly, subtracting decreases. In math lingo, it may decrease or increase or neither.

In English, "adding" and "subtracting" are opposite. In math lingo, they're opposite, and yet they're the same! For adding a number is the same as subtracting some other number (its negative).

In English, "multiplying" means repeated adding. It makes things bigger. In math lingo, multiplying makes them bigger, smaller, or neither, depending on what you multiply with.

Correspondingly, "divide" means cut into pieces, possibly equal pieces. In math lingo, "divide" is the same as "multiply," in the sense that dividing by a number other than zero is the same as multiplying by some other number (its reciprocal).

There's a familiar conundrum about amoebas: *amoebas multiply by dividing.* To untangle this nonsensical but correct statement, you must see the difference between the mathematical and the plain English meanings of "multiply" and "divide."

What should you do about all this? Be aware of it and point it out to students. By appropriate examples, make them realize that what they hear in class or read in the text is technical jargon, not plain English. Otherwise, when they try to remember what you said in yesterday's lecture, they may remember it with the wrong meaning (the plain English).

Anneli Lax reminded me of one of the commonest linguistic pitfalls: the little one-letter word "a." Her example is "Show that a number divisible by 6 is even."

No seasoned math teacher is surprised to receive the wrong answer, "42 is divisible by 6. 42 is even." Why is this answer wrong? 42 *is* divisible by 6, and 42 *is* even. What's wrong is that the question has been misunderstood. By "a," the questioner meant "every"; the student misinterpreted it as "some." This is a quantification problem, which in principle could be cured by using symbolic logic instead of English. But in a case like this, something deeper is wrong. The student should realize that with the interpretation "some," the question is too trivial to be on the test. Grounding in the context saves the student from most verbal pitfalls. One goal of teaching is to ground the student in the context. Linguistic ambiguities can hurt.

In logic, the pitfalls of "or" and "implies" are familiar.

Take "or." In plain English, "Tea or coffee?" means one or the other, not both. It's called the "exclusive or."

"Are you coming or going?"

"Was that your husband or your boy friend?"

"Do it now or later?"

All are exclusive. It's hard to think of a colloquial example of the other "or," the inclusive one. A reasonable example might be, "Like a hug or a kiss?"

In logic, "or" is inclusive by convention. "A or B" is true if A or B *or both* is the case. I think it's customary to explain on the first day of elementary logic class that logicians have decreed "or" to be inclusive. A student can accept that logicians felt they had to pick one or the other. Perhaps they had a reason for picking the inclusive.

Peter Lax tells about the famous logician Abraham Fraenkel, of German origin and Israeli residence. Once in Jerusalem or Tel Aviv he was on a bus scheduled to leave the station at 9A.M. At 9:05 the bus was still sitting in the station. Fraenkel waved a bus schedule at the bus driver, who asked, "What are you, a German or a professor?" Fraenkel inquired in return, "Do you use the inclusive 'or' or the exclusive?"

"Implies" is worse. In plain English, "A implies B" means that if A is true, B must be true. If A is false, the "implies" statement is vacuous, neither true nor false.

But in logic, the "law of the excluded middle" insists that every statement be either true or false. The statement "A implies B" has to be either true or false, even if A is false. Logicians chose "true." So in logic, if A is false, then A implies B, *whatever B may be*. This is so unintuitive, I say logicians should have used another word, even made up a word. It's too late for that. But the student is told that "implies" in logic is different from "implies" in plain English. In pre-calculus, calculus, and post-calculus, we should be equally considerate to warn of linguistic traps.

I have just carelessly used "equally." "Equal" is used freely, from kindergarten to postgraduate. It's never defined or explained.

In plain English, its meaning varies. Sometimes it's "identical, indistinguishable." Sometimes it's "worth the same number of dollars." Or "just as good" for some purpose.

Math lingo sometimes says "equal," sometimes "equivalent," the latter if an equivalence relation has been defined. Then we explain that an equivalence relation is Reflexive, Symmetric, and Transitive; it defines a partition on a set.

But what does *equal* mean? When we say $1/2 = 2/4$, we don't mean $1/2$ is indistinguishable from $2/4$. They have different numerators. They have different denominators. We regard them as *equivalent* for good and sufficient reasons. All this may be explained in an advanced course, on the rare occasion when a detailed construction of the rationals is carried out. But already in the fourth grade the $=$ relation is an equivalence relation between fractions, not an identity. No one ever explains this, so there's no way for the student to understand $=$, except in terms of models like slices of apple pie.

This nonunderstanding was manifested frighteningly when a calculus student was asked, "What is the minimum of the function

$$y = x^2 + 2x + 5?"$$

and answered "correctly"

$$\text{“}x^2 + 2x + 5 = 2x + 2 = -1 = 4 \textit{ minimum}\text{”}$$

Maybe this is the outcome of years in high school spent factoring, multiplying, and dividing expressions that always remained equal.

In plain English, set and group are synonyms. When we teach groups, we define set and group, then charge ahead. But some students wonder, "What's the difference? A group is the same as a set." Mention this plain English equivalence, and state explicitly that in math these words have different meanings.

The same is true of sequence and series. Their plain English meanings are the same—what in math lingo we call "a finite list." "Series" is more colloquial than sequence—for example, it's the World Series, not the World Sequence! Here the danger of confusion is more serious than with set and group. The mathematical meanings of sequence and series are so close that the distinction between them is crucial. In teaching series, we should acknowledge that we're giving a new meaning to a common word: putting + signs instead of commas between the terms.

The first day of first-semester calculus I like to talk about driving to Santa Fe. Distance from Albuquerque is a function of time. Speed is another function of time. But what is "function" in English? If you ask, "Of what is the speed a function?," you're told, "It's a function of how much gas you give" or "Of how hard you push the accelerator pedal." "Function" in English (apart from the irrelevant reference to weddings and Bar Mitzvahs) involves causal dependence. "How fast you learn is a function of how hard you study," for example. How can *anything* be a function of *time*? But the students swallow that. They understand a graph with a time axis. Then I say, "Distance is a monotonic increasing function of time, so the inverse function exists. Time is a function of distance." How can time, the independent, uncaused variable, be *caused* by *distance*? We try to teach our technical meaning of "function" without noticing the meaning the student brings into class.

We're aware that "limit" and "converge" are deep concepts. We sweat over them. But we don't acknowledge the complication caused by plain English. A

"limit" in English is a barrier, a boundary beyond which one may not pass. This may partly explain why students want to approach a limit only from one side, not in alternating fashion. As for "converge." In practical computation, an algorithm converges when it settles down to one value and stays there—stays till whoever's doing the calculation is satisfied. That's the English of converge—"settle down" "close to" some "limit." In teaching our uncomputational, abstract meaning of "converge," we should talk about the colloquial meaning and explain the difference.

In advanced mathematics, there's more linguistic confusion. Surds (absurd), irrational and imaginary numbers, singular perturbations, degenerate kernels, strange attractors—all sound dangerous, undesirable, things to avoid. Yet a degenerate kernel or a singular perturbation may be more useful than a non-degenerate or regular one.

We also talk about "function spaces." The points in a function space are functions. But a function is a graph—a curve. How can a *curve* be a *point*? A point, which has *no parts*! We don't acknowledge the change of meaning. Just give a definition and two examples, then charge ahead.

An example of the opposite kind (due to Peter Lax) is "simple curve." Draw a confusing tangle that doesn't intersect itself. It's complicated. We say it's simple.

What about "partial?" A partial order isn't a special kind of order. A partial differential equation isn't part of an ordinary differential equation. And an ordinary differential equation may well be extraordinary.

Exercise: (a) give the plain English meaning of prime; differentiate; integrate.

(b) check your answers against a standard dictionary.

(c) make up three slogans, one using each of these three words, that could appear on picket signs at a demonstration.

It's fortunate that some double meanings are so far apart they can be used for a joke. A manifold is part of an automobile engine (I think), and a commutator is part of a direct current electric motor.

ACKNOWLEDGMENTS. Veronka John-Steiner, Anneli Lax, and Peter Lax gave suggestions and encouragement. [1] is an inspiring example of frank talk about college math teaching.

REFERENCE

[1] Boas, R. P. Can we make mathematics intelligible?, *Amer. Math. Monthly*, 88 (1981), 727–731.

1000 Camino Rancheros
Santa Fe, N.M. 87501
rhersh@math.unm.edu

NOTES

Edited by **Jimmie D. Lawson**

Three Secrets About Harmonic Functions

R. B. Burckel

1. INTRODUCTION AND NOTATION. I'll not be coy, but reveal these to the reader at once. Actually, they're not secrets to practitioners, but only to the broader mathematical public. Since they are striking results and accessible by elementary means, perhaps their revelation will be welcomed. They say roughly:

1. There exists a simply connected plane region for which the Dirichlet problem is not solvable.
2. The areal mean-value property of harmonic functions characterizes disks.
3. Harmonicity of the function that measures the distance to the boundary of a region characterizes half-planes.

In the following sections I will give the precise statements and proofs, but first some necessary notation. Let $D(a,r) := \{z \in \mathbb{C}: |z-a| < r\}$ for $a \in \mathbb{C}$, $r \geq 0$ and $\mathbb{D} := D(0,1)$. $\mathbb{H}$ denotes the open right half-plane $\{z \in \mathbb{C}: \operatorname{Re} z > 0\}$. For $a, b \in \mathbb{C}$ let $[a,b]$ denote the interval $\{(1-t)a + tb: 0 \leq t \leq 1\}$. $\mathbb{C}_\infty$ is the one-point compactification of $\mathbb{C}$ (a.k.a. the Riemann sphere). A *region* is a non-void open, connected subset Ω of $\mathbb{C}$. We write $\partial\Omega := \overline{\Omega} \setminus \Omega$, where bar denotes closure in $\mathbb{C}$. If Ω is unbounded, we write $\partial_\infty\Omega := \{\infty\} \cup \partial\Omega$. Of the myriad equivalent notions of simple connectivity in $\mathbb{C}$ we adopt this one: Ω *simply connected* means that $\mathbb{C}_\infty \setminus \Omega$ is connected, or what is the same thing, $\mathbb{C} \setminus \Omega$ has no bounded component.

2. A COUNTEREXAMPLE. We say that the *Dirichlet problem is solvable in* Ω if every continuous $f\colon \partial\Omega \to \mathbb{R}$ admits a continuous extension $F\colon \overline{\Omega} \to \mathbb{R}$ such that F is harmonic in Ω. If f is bounded, the Perron-Kellogg-Wiener construction produces a harmonic function $F\colon \Omega \to \mathbb{R}$ that satisfies

$$\lim_{\substack{z \to x \\ z \in \Omega}} F(z) = f(x)$$

for every $x \in \partial\Omega$ at which Ω possesses a *barrier*. Every simply connected region in $\mathbb{C}$ has a barrier at every one of its boundary points. (For these facts see **[4]**.) The continuous function f is certainly bounded if $\partial\Omega$ is bounded, or if $\partial\Omega$ is unbounded and f is actually continuous on the (compact) set $\partial_\infty\Omega$, i.e., $\lim_{\substack{z \to \infty \\ z \in \Omega}} f(z)$ exists in $\mathbb{R}$.

In view of all this, Theorem 1 might come as a shock. As far as I know, such examples are only of recent origin; but see **[3]**. Specifically, this one originated in **[6]**. After the author presented a somewhat simplified version to a seminar,

Sadahiro Saeki introduced further simplifications. Here is his version

Theorem 1. *There exists a simply connected plane region for which the Dirichlet problem is not solvable.*

Proof: For each integer $n \geq 3$ define the open sector

$$S_n := \left\{ re^{i\theta} : 0 < r < n, \left| \theta - \frac{\pi}{n} \right| < \frac{\pi}{2n(n+1)} \right\}.$$

Since the restriction on θ here is

$$\frac{1 + \dfrac{1}{2n}}{n+1} < \frac{\theta}{\pi} < \frac{1 + \dfrac{1}{2(n+1)}}{n},$$

it is obvious that

$$\bar{S}_m \cap \bar{S}_n = \{0\} \qquad \text{whenever } m > n \geq 3. \tag{1}$$

Define

$$\Omega := D(0,2) \cup \bigcup_{n \geq 3} S_n.$$

This set is open and is starlike with respect to 0; hence it is connected. Hence too $\mathbb{C} \setminus \Omega$ is a union of half-lines, so $\mathbb{C} \setminus \Omega$ has no bounded component, making Ω simply connected. Notice that $\partial\Omega$ consists of $[2, +\infty[$, together with arcs of $\partial D(0,2)$, together with the disjoint sets $(\partial S_n) \setminus D(0,2)$ for $n \geq 3$. Now it is clear that $z \mapsto (ze^{-i\pi/n})^{n(n+1)}$ maps S_n into $\mathbb{H}$ and the boundary rays of S_n into $i\mathbb{R}$. From this and (2) it is easy to see that a function $f\colon \partial\Omega \to \mathbb{R}$ is well defined by

$$f(z) := \begin{cases} \operatorname{Re}\left[\left(ze^{-i\pi/n} \right)^{n(n+1)} \right] & \text{for } z \in (\partial S_n) \setminus D(0,2),\ n \geq 3 \\ 0 & \text{for all other } z \in \partial\Omega \end{cases}$$

and is continuous. Suppose h is a continuous real-valued extension of f to $\bar{\Omega}$ that is harmonic in Ω. On the compact subset $\bar{D}(0,2) \subset \bar{\Omega}$, h is bounded below, say by $c \in]-\infty, 0[$. For each integer $n \geq 3$

$$h(z) - \operatorname{Re}\left[\left(ze^{-i\pi/n} \right)^{n(n+1)} \right] \geq c \qquad \text{for all } z \in \partial S_n. \tag{2}$$

This is because for $z \in (\partial S_n) \setminus D(0,2) \subset \partial\Omega$, the left-hand side of (2) is $h(z) - f(z) = 0 > c$, while for $z \in (\partial S_n) \cap D(0,2)$ it is $h(z) - 0 \geq c$, by definition of c. By the Maximum Principle [**4**, p. 253], inequality (2) must hold as well at $z = 2e^{i\pi/n} \in S_n$, yielding

$$h(2e^{i\pi/n}) - 2^{n(n+1)} \geq c. \tag{3}$$

The validity of (3) for all $n \geq 3$ makes h unbounded on the compact subset $\bar{D}(0,2)$ of $\bar{\Omega}$. This contradiction shows that no such function h exists. ■

3. A CHARACTERIZATION OF DISKS. The second secret is surprising because it's an analytic characterization of a geometric property. Recall the *areal mean-value*

property of harmonic functions:

(AM) $\quad U$ open $\subset \mathbb{C}$, $\overline{D}(w,R) \subset U$, $h\colon U \to \mathbb{R}$ harmonic $\Rightarrow$

$$h(w) = \frac{1}{\pi R^2}\int_{D(w,R)} h(z)\, d\lambda(z),$$

where λ is two-dimensional Lebesgue (area) measure. Using Lebesgue's Monotone Convergence Theorem we can slightly extend the scope of (AM): its conclusion holds whenever

$$D(w,R) \subset U \text{ and } \int_{D(w,R)} |h(z)| d\lambda(z) < +\infty.$$

In this form there is an unexpected converse:

Theorem 2. *Suppose U open $\subset \mathbb{C}$ has finite area, $z_0 \in U$, and*

$$(*) \qquad h(z_0) = \frac{1}{\lambda(U)}\int_U h(z)\, d\lambda(z)$$

for every h that is harmonic in and (absolutely) *integrable over U. Then U is a disk with center z_0.*

Proof: Let r be the radius of the largest open disk, call it D, centered at z_0, which lies in U. This number is finite and positive and there exists (compactness) some $z_1 \in \mathbb{C} \setminus U$ with $|z_1 - z_0| = r$. We will show that $D = U$. To this end, define the function h on U by

$$h(z) := \frac{|z - z_0|^2 - r^2}{|z - z_1|^2} + 1, \qquad z \in U. \tag{4}$$

Since $|z_1 - z_0| = r$, a little computation reveals that

$$h(z) = 2\,\mathrm{Re}\left(\frac{z - z_0}{z - z_1}\right), \tag{5}$$

so h is harmonic in U. To check that h is integrable over U, (5) shows that it suffices, since U has finite area and h is bounded outside any neighborhood of z_1, to show that $|z - z_1|^{-1}$ is integrable over $D(z_1, 1)$. That follows from passage to polar coordinates:

$$\int_{D(z_1,1)} |z - z_1|^{-1}\, d\lambda(z) = \int_{D(0,1)} |z|^{-1}\, d\lambda(z) = \int_0^1\int_0^{2\pi} |\rho e^{i\theta}|^{-1}\rho\, d\theta\, d\rho = 2\pi.$$

Now $(*)$ ensures that

$$0 = h(z_0) = \int_U h\, d\lambda = \int_D h\, d\lambda + \int_{U\setminus D} h\, d\lambda. \tag{6}$$

But according to (AM) we also have

$$0 = h(z_0) = \int_D h\, d\lambda.$$

Combined with (6) and the fact $h \geq 1$ in $U \setminus D$ (see (4)), this gives

$$0 = \int_{U\setminus D} h\, d\lambda \geq \lambda(U \setminus D) \geq \lambda(U \setminus \overline{D}).$$

It follows that the open set $U \setminus \overline{D}$ must be empty, that is, $U \subset \overline{D}$. From $D \subset U$ open $\subset \overline{D}$ it follows that $U = D$. ∎

This proof is due to Kuran [7], and represents the final touch of elegance and hypothesis-reduction in a long evolution. In fact, this result is part of a large subject called quadrature problems that interested readers can find more about in [9]. The hypothesis of Theorem 2 can be weakened to require only that $(*)$ hold for all nonnegative harmonic functions h that are integrable over U. However, it is not sufficient that $(*)$ hold only for bounded harmonic functions h; see [1].

4. ANALYTIC CHARACTERIZATION OF HALF-PLANES. The third secret is also an analytic characterization of a geometric property. Theorem 3 occurs as an exercise in the little book of Fuchs and Schumitsky [5], but in a letter to me Professor Fuchs indicated that he is unaware of its provenance. Armitage and Kuran [1] investigate similar problems and prove that the function ρ in (7) is subharmonic in Ω if and only if $\mathbb{C} \setminus \Omega$ is convex. Using their result, Parker [8] gave a proof of Theorem 3 somewhat different from that presented here. The proof that follows was helped to birth by conversations with Erich Novak and correspondence with Wolfhard Hansen. For a region Ω that is not all of $\mathbb{C}$ let

$$\rho(z) := \inf\{|z - w| : w \in \mathbb{C} \setminus \Omega\}, \qquad z \in \mathbb{C}, \tag{7}$$

be the function "distance to the complement of Ω." It is well known and elementary that ρ is a continuous function on $\mathbb{C}$, even a contractive mapping, whose zero-set is $\mathbb{C} \setminus \Omega$. Each function $z \mapsto |z - w|$ is subharmonic in $\mathbb{C}$, although this fact plays no role in the sequel.

Theorem 3. *If ρ is harmonic in Ω, then Ω is a half-plane.*

Of course, the converse of this is trivially also true. For the proof of the theorem we need an elementary but perhaps not so well known fact about harmonic functions. A direct proof is indicated, although it can also be deduced from better-known properties of holomorphic functions by symmetry arguments.

Lemma. *Suppose $a, b \in \mathbb{C}$, $[a, b] \subset U$ open $\subset \mathbb{C}$, $h \colon U \to \mathbb{R}$ harmonic. If $h^{-1}(0) \cap [a, b]$ is infinite, then $h[a, b] = 0$.*

Proof: We may suppose $[a, b] \subset \mathbb{R}$. Let x_0 be a limit point of $h^{-1}(0) \cap [a, b]$. For some $r > 0$, $D(x_0, r) \subset U$. Harmonic functions are real-analytic, so $h(x) = h(x + i \cdot 0)$ is represented by a power series in $x - x_0$ for $x \in]x_0 - r, x_0 + r[$. The usual argument that the zeros of a non-trivial power series are isolated [4, p. 78] can be applied here to show that $h = 0$ in this whole interval. This shows that the closed set of all limit points of $h^{-1}(0) \cap [a, b]$ is also relatively open in $[a, b]$. As this set is non-empty by hypothesis, it must be all of the connected set $[a, b]$. ∎

Proof of Theorem 3. By definition of ρ

$$\Omega = \rho^{-1}(]0, +\infty[) \qquad \text{and} \qquad D(z, \rho(z)) \subset \Omega \qquad \text{for all } z \in \Omega. \tag{8}$$

Fix $x_0 \in \Omega$. By compactness there exists $b_0 \in \partial\Omega$ with $\rho(x_0) = |x_0 - b_0|$. After translation and rotation of Ω we can assume

$$b_0 = 0, \quad x_0 \text{ is real and positive.}$$

Then an application of the triangle inequality shows that $\rho(x) = x$ for all $x \in]c, x_0]$:

$$D(x_0, x_0) \subset \Omega \qquad \text{and} \qquad \rho(x) = x \qquad \forall x \in]0, x_0]. \tag{9}$$

From the lemma it follows that $\rho(x) = x$ for all $x \in]0, 2x_0[\subset D(x_0, x_0)$, whence by continuity of ρ on $\mathbb{C}$, $\rho(2x_0) = 2x_0 > 0$. It follows from (8) that

$$D(2x_0, 2x_0) \subset \Omega \qquad \text{and} \qquad \rho(x) = x \qquad \forall x \in]0, 2x_0]. \tag{10}$$

Iterating the argument that led from (9) to (10) shows that

$$\mathbb{H} = \bigcup_{n \in \mathbb{N}} D(2^n x_0, 2^n x_0) \subset \Omega \qquad \text{and} \qquad \rho(x) = x \qquad \forall x \in]0, +\infty[. \tag{11}$$

Consequently, for $z \in \mathbb{H}$

$$\rho(z) = \operatorname{dist}\{z, \mathbb{C} \setminus \Omega\} \geq \operatorname{dist}\{z, \mathbb{C} \setminus \mathbb{H}\} = \operatorname{Re} z. \tag{12}$$

Now (11) and (12) say that $\rho(z) - \operatorname{Re} z$ is harmonic and non-negative in $\mathbb{H}$, but vanishes at (many) points of $\mathbb{H}$, so by the Maximum Principle it vanishes identically in $\mathbb{H}$:

$$\rho(z) = \operatorname{Re} z \qquad \forall z \in \mathbb{H} \subset \Omega. \tag{13}$$

Since ρ is continuous on $\mathbb{C}$, this shows that $\rho(i\mathbb{R}) = 0$, so by (8), $i\mathbb{R} \cap \Omega = \varnothing$. Being connected, Ω cannot therefore meet both $\mathbb{H}$ and $\mathbb{C} \setminus \overline{\mathbb{H}}$, so from (13) it follows that

$$\mathbb{H} = \Omega.$$

ACKNOWLEDGMENTS. I thank the referee for spotting some deficiencies in the first draft, and David Armitage for some of the history in Section 4.

REFERENCES

1. D. H. Armitage and M. Goldstein, The volume mean-value property of harmonic functions, *Complex Variables Theory and Application* 13 (1990), 185–193. MR 91a:31004.
2. D. H. Armitage and Ü. Kuran, The convexity of a domain and the superharmonicity of the signed distance function, *Proc. Amer. Math. Soc.* 93 (1985), 598–600. MR 86k:31005.
3. M. Brelot, Sur la mesure harmonique et le problème de Dirichlet, *Bulletin des Sciences Mathématiques* (2) 69 (1945), 153–156. MR 7, p. 522.
4. J. B. Conway, *Functions of One Complex Variable*, Graduate Texts in Mathematics, vol. 11, Springer-Verlag (2nd ed., 1978), New York. MR 80c:30003.
5. W. H. J. Fuchs, *Topics in the Theory of Functions of One Complex Variable*, Manuscript prepared with the collaboration of Alan Schumitsky. Van Nostrand Mathematical Studies #12, D. Van Nostrand Co. (1967), Princeton. MR 36.3954.
6. S. J. Gardiner, The Dirichlet problem with non-compact boundary, *Math. Zeit.* 213 (1993), 163–170, MR 94e:31001.
7. Ü, Kuran, On the mean-value property of harmonic functions, *Bull. London Math. Soc.* 4 (1972), 311–312. MR 47.8887 and 52.8464.
8. M. J. Parker, Mean values and distance functions in potential theory, Ph.D. thesis (1987), University of Liverpool.
9. M. Sakai, *Quadrature Domains*, Lecture Notes in Mathematics 934, Springer-Verlag (1982), New York. MR 84h:41047.

Department of Mathematics
Kansas State University
Manhattan, KS 66506

A Short Proof of the Erdős-Mordell Theorem

Vilmos Komornik

Dedicated to the memory of P. Erdős (1913–1996)

The following beautiful inequality was conjectured by Erdős in [**1**]:

Theorem. *Given a point P inside a triangle ABC, let us denote by R_a, R_b, R_c its distance from the vertices A, B, C and by r_a, r_b, r_c its distance from the lines of the sides $a = BC$, $b = CA$ and $c = AB$; see Figure* 1. *Then*

$$R_a + R_b + R_c \geq 2(r_a + r_b + r_c). \tag{1}$$

He arrived at his conjecture in an experimental way in 1932, after having drawn many triangles.

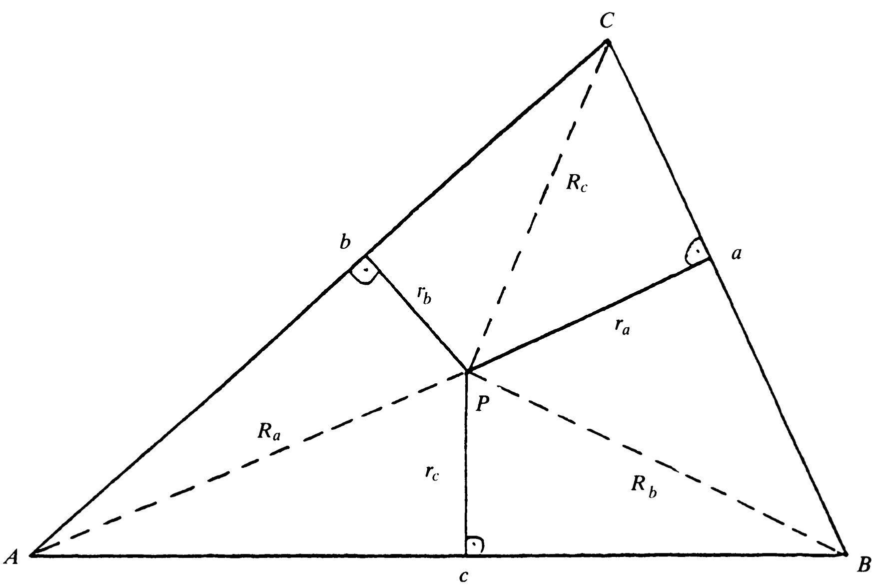


Figure 1

The first proofs by Mordell [**2**], [**3**] and Barrow [**3**] were based on trigonometry. Later, several elementary proofs were given, using either not too well-known results (a theorem of Pappus in [**4**], a theorem of Ptolemy in [**7**]) or clever angular computations with similar triangles in [**5**]. In [**6**] a whole series of related inequalities was established by elementary means; for the proof of (1), however, a nontrivial transformation (using isogonal conjugates) was applied.

The purpose of this note is to give still another elementary proof. We use only basic notions and results taught in secondary (or even elementary) schools. At the same time, our proof seems to be shorter and simpler than any previous one. As usual, we also show that the inequality in (1) is strict unless the triangle is equilateral and P is its center.

At the end of this note we recall briefly how to deduce several other well-known geometric inequalities from this theorem. More applications are given in **[8]**.

1. Consider first a point P lying on the side $a = BC$ (see Figure 2). Then the double area of the triangle APC is equal to br_b, the double areas of the triangle ABP is equal to cr_c and hence that of the triangle ABC is $br_b + cr_c$. On the other hand, R_a cannot be shorter than the altitude of ABC drawn from A. Therefore

$$aR_a \geq br_b + cr_c. \tag{2}$$

Observe that this inequality remains valid for every point P in the angular domain BAC: it suffices to note that inequality (2) is equivalent by similarity to the inequality corresponding to the intersection P' of the side BC with the ray AP.

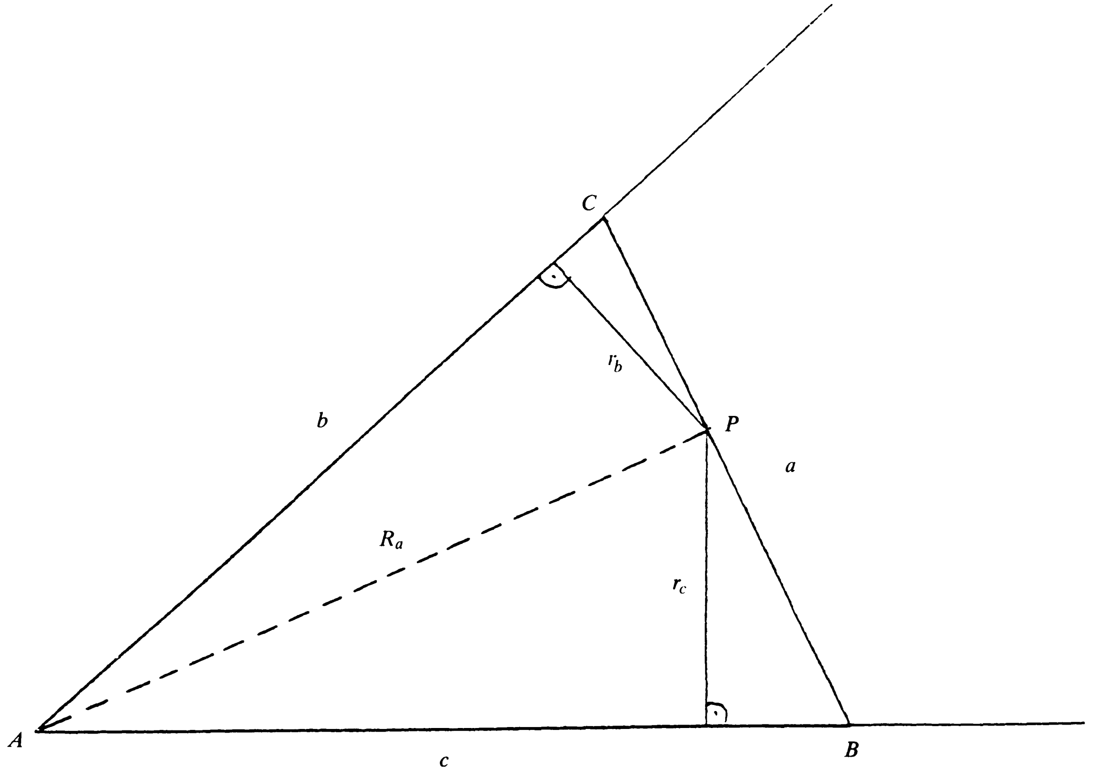


Figure 2

2. If the triangle is equilateral, then the inequality (1) follows at once. Indeed, it is sufficient to divide the inequality (2) by $a = b = c$ and add it to the two analogous inequalities $R_b \geq r_c + r_a$, $R_c \geq r_a + r_b$, obtained by cyclical permutation of the indices.

Moreover, the proof of (2) shows that we have equality in (1) only if P belongs to all three altitudes of the triangle, i.e., if P is the center of the equilateral triangle ABC.

3. If the triangle is not equilateral, then we need a simple corollary of the inequality (2). Applying (2) to the reflection P' of P on the bisector of the angle BAC, we have $R'_a = R_a$, $r'_b = r_c$ and $r'_c = r_b$ with obvious notation. Hence

$$aR_a \geq br_c + cr_b$$

for any point P in the angular domain BAC.

If P is inside the triangle ABC, then we also have $bR_b \geq cr_a + ar_c$ and $cR_c \geq ar_b + br_a$ by symmetry. The inequality (1) now follows easily:

$$R_a + R_b + R_c \geq \frac{b^2 + c^2}{bc} r_a + \frac{c^2 + a^2}{ca} r_b + \frac{a^2 + b^2}{ab} r_c > 2(r_a + r_b + r_c).$$

Note that we have strict inequality in the last step because a, b, and c are not all equal. ■

Now let us give some applications. In what follows we shall denote by m_a, m_b, m_c, R, r, and T the altitudes, the circumradius, the inradius, and the area of the triangle ABC, respectively. Note that for any point P inside the triangle we have clearly

$$ar_a + br_b + cr_c = 2T; \tag{3}$$

applying this to the incenter we obtain the well-known equality

$$(a + b + c)r = 2T. \tag{4}$$

(a) For any point P inside the triangle we have

$$R_a + r_a \geq m_a, \qquad R_b + r_b \geq m_b, \qquad \text{and} \qquad R_c + r_c \geq m_c. \tag{5}$$

Summing them and applying the inequality (1) we find that

$$m_a + m_b + m_c \leq 1.5(R_a + R_b + R_c). \tag{6}$$

By continuity, this inequality remains valid if P is on the boundary of the triangle. Moreover, it remains valid for all points P in the plane. Indeed, if P is outside of the triangle, then the sum $R_a + R_b + R_c$ decreases as we replace P by its orthogonal projection on the (closed convex) triangle ABC.

(b) Since

$$\frac{m_a + m_b + m_c}{3} \geq \frac{3}{m_a^{-1} + m_b^{-1} + m_c^{-1}} = \frac{6T}{a + b + c} = 3r \tag{7}$$

by (4), we conclude from (6) that

$$R_a + R_b + R_c \geq 6r. \tag{8}$$

(c) Applying (6) and (7) to the circumcenter (so that $R_a = R_b = R_c = R$) we obtain

$$9r \leq m_a + m_b + m_c \leq 4.5R. \tag{9}$$

In particular, this implies the well-known inequality $R \geq 2r$. Note that this last inequality also follows at once from (8) applied to the circumcenter.

(d) Let us give finally a direct proof (in the same spirit) of the inequality $R \geq 2r$, without using the Erdős-Mordell theorem. It follows from (5) that

$$aR_a + bR_b + cR_c + ar_a + br_b + cr_c \geq am_a + bm_b + cm_c = 6T.$$

Applying (3) and (4) we conclude that

$$\frac{aR_a + bR_b + cR_c}{a + b + c} \geq 2r. \tag{10}$$

Like (6), this inequality also remains valid even if P is outside the triangle. Applying it to the circumcenter, the inequality $R \geq 2r$ follows.

REFERENCES

1. P. Erdős, Problem 3740, this *Monthly* 42 (1935), 396.
2. L. J. Mordell, Egy geometriai probléma megoldása (Solution of a geometrical problem), *Középiskolai Matematikai és Fizikai Lapok* 11 (1935), 145–146.
3. Solution of Problem 3740, this *Monthly* 44 (1937), 252–254. (Solutions by L. J. Mordell and by D. F. Barrow.)
4. D. K. Kazarinoff, A simple proof of the Erdős-Mordell inequality for triangles, *Michigan Math. J.* 4 (1957), 97–98.
5. L. Bankoff, An elementary proof of the Erdős-Mordell theorem, this *Monthly* 65 (1958), 521.
6. A. Oppenheim, The Erdős inequality and other inequalities for a triangle, this *Monthly* 68 (1961), 226–230.
7. A. Avez, A short proof of a theorem of Erdős and Mordell, this *Monthly* 100 (1993), 60–62.
8. Reiman István, 60 éve jelent meg a Középiskolai Matematikai Lapokban az Erdős-Mordell tétel (On the sixtieth anniversary of the publication of the Erdős-Mordell theorem in this journal), *Középiskolai Matematikai és Fizikai Lapok* 45 (1995), 385–394.

Institut de Recherche Mathématique Avancée,
Université Louis Pasteur et C.N.R.S.,
7, rue René Descartes,
67084 Strasbourg Cédex, France
komornik@math.u-strasbg.fr

Inverting the Difference of Hilbert Space Projections

Don Buckholtz

Let R and K be subspaces of a Hilbert space H, and let P_R and P_K denote the orthogonal projections of H onto these subspaces. When is the operator $P_R - P_K$ invertible? We show here that the obvious necessary condition, $H = R \oplus K$, is sufficient as well. We also find the inverse.

Theorem. *Let R and K be subspaces of a Hilbert space H, and let P_R and P_K denote the orthogonal projections of H onto these subspaces. The following are equivalent:*

(i) *The operator $P_R - P_K$ is invertible.*
(ii) *H is the direct sum of R and K.*
(iii) *There exists a linear idempotent M with range R and kernel K.*

If $P_R - P_K$ is invertible, then $(P_R - P_K)^{-1} = M + M^* - I$.

Proof: The equivalence of (ii) and (iii) is well known and easy to prove. What needs to be shown is that (i) and (iii) are equivalent. Suppose first that there exists an idempotent M with range R and kernel K. We have $MP_K = 0$ and $P_R M = M$. Since $I - M$ is idempotent with range K and kernel R, we have the corresponding results $(I - M)P_R = 0$ and $P_K(I - M) = I - M$. Therefore

$$(M + M^* - I)(P_R - P_K) = (P_R M + P_K(I - M))^* - (I - M)P_R - MP_K$$
$$= (M + I - M)^* = I.$$

Taking the adjoint yields $(P_R - P_K)(M + M^* - I) = I$; therefore, $(P_R - P_K)^{-1} = M + M^* - I$.

It remains to show that the invertibility of $P_R - P_K$ implies the existence of an idempotent M with range R and kernel K. To obtain this result, premultiply and postmultiply the identity

$$(P_R - P_K)P_R = (I - P_K)(P_R - P_K)$$

by $(P_R - P_K)^{-1}$ and call the resulting operator M. We shall show that

$$M = P_R(P_R - P_K)^{-1} = (P_R - P_K)^{-1}(I - P_K)$$

is an idempotent with range R and kernel K. That M has range R is a consequence of the first expression for M; from the second expression it follows that M has kernel K.

To establish that M is idempotent, note that

$$M - I = P_R(P_R - P_K)^{-1} - (P_R - P_K)(P_R - P_K)^{-1} = P_K(P_R - P_K)^{-1}.$$

Using the fact that $(I - P_K)P_K = 0$, we obtain

$$M^2 - M = M(M - I) = (P_R - P_K)^{-1}(I - P_K)P_K(P_R - P_K)^{-1} = 0.$$

REFERENCES

1. P. R. Halmos, *Finite-Dimensional Vector Spaces*, Springer-Verlag, New York, 1974.
2. P. R. Halmos, *A Hilbert Space Problem Book*, 2nd ed., Springer-Verlag, New York, 1982.

The University of Kentucky
Lexington, Kentucky 40506

Problems from 100 years Ago in the MONTHLY ...

70. Proposed by J. A. CALDERHEAD, M.Sc., Professor of Mathematics in Curry University, Pittsburgh, Pennsylvania

A owes me \$100 due in 2 years, and I owe him \$200 due in 4 years; when can I pay him \$100 to settle the account equitably, money being worth 6%? [p. 12]

79. Proposed by F. M. PRIEST, St. Louis, Missouri.

How many \$20 gold pieces can be put in a room 20 feet long, 18 feet wide, and 9 feet high? [p. 121]

56. Proposed by H. C. WHITAKER, A.M., Ph.D., Professor of Mathematics, Manual Training School, Philadelphia, Pennsylvania.

"Hey-diddle-diddle, the cat and the fiddle,
The cow jumped over the moon."

Taking the weight of the cow to be 600 pounds, the initial resistance of the air to be 100 pounds and varying as the square of the velocity, find the initial and final velocities and the times of rising and falling. [p. 124]

... vol. 4, 1897.

THE EVOLUTION OF . . .

Edited by **Abe Shenitzer**

Mathematics, York University, North York, Ontario M3J 1P3, Canada

Glimpses of Algebraic Geometry[1]

I. G. Bashmakova and E. I. Slavutin

PLANE ALGEBRAIC CURVES. Consider the equation

$$F(x,y) = 0, \tag{1}$$

where $F(x, y)$ is a polynomial with rational coefficients that is irreducible over the field $\mathbb{Q}$ of rational numbers. The set of points of the real plane $\mathbb{R}^2$ whose coordinates satisfy the equation (1) is called a *plane (rational) algebraic curve*. If F is linear then we speak of a *rational line*. The points with rational coordinates are called *rational points*.

By the *order* of the curve Γ defined by equation (1) we mean the degree n of the polynomial $F(x, y)$. The number of points of intersection of Γ and an arbitrary line $Ax + By + C = 0$ is exactly n. When counting the number of points of intersection we must consider multiplicities, complex points, and points at infinity. We give a few illustrative examples.

a. The curve $x^2 + y^2 = 1$ and the straight line $x + y = 10$ intersect in two complex points;
b. The curve $y^3 = 1 - x^3$ and the straight line $y = 1$ have the triple point of intersection $P(0, 1)$;

(*Remark.* For a discussion of singular and multiple points see [1] and [5]. (Trans.))

c. The curve $y^2 = 4x^2 + x + 2$ and the straight line $y = 2x$ have two points of intersection, namely the point $M(-2, -4)$ and a point at infinity.

In order to define points at infinity we must introduce homogeneous coordinates, that is, essentially, we must go from the real plane $\mathbb{R}^2$ to the projective plane $\mathbb{P}^2$. A point of the projective plane is given by an ordered triple of real numbers (u, v, w) not all of which are 0. Proportional triples define the same point. The numbers in a triple (u, v, w) are called *homogeneous coordinates in* $\mathbb{P}^2$.

We now determine a (partial) correspondence between the points on $\mathbb{R}^2$ and on $\mathbb{P}^2$. Let (u, v, w) be a point on $\mathbb{P}^2$. If $w \neq 0$, then $(u/w, v/w, 1)$ determines the same point on $\mathbb{P}^2$. We associate with it the point on $\mathbb{R}^2$ with coordinates $x = u/w$, $y = v/w$. If $w = 0$, then the point $(u, v, 0)$ has no "partner" on $\mathbb{R}^2$. We call such points *points at infinity*. All points at infinity lie on the line at infinity $w = 0$.

[1] This is a translation (by Abe Shenitzer, who also titled the piece) of part of the introduction to the monograph by Bashmakova and Slavutin titled *A History of Diophantine Analysis from Diophantus to Fermat*, published in 1984.

In order to change equation (1) to an equation in homogeneous coordinates we put $x = u/w$, $y = v/w$. After obvious simplifications we obtain a homogeneous equation of the form

$$\Phi(u, v, w) = 0. \tag{2}$$

Now points at infinity have the same status as ordinary points.

In terms of homogeneous coordinates, our curve $y^2 = 4x^2 + x + 2$ has the equation $v^2 = 4u^2 + uw + w^2$. Putting $w = 0$, we obtain its two rational points at infinity $M_1(1, 2, 0)$ and $M_2(1\ -2, 0)$. The line $v = 2u$ passes through the point M_1. This is its second point of intersection with our curve.

The classification of curves by order is of great significance. It was introduced by Descartes (who put in the same class curves of order $2n$ and $2n - 1$) and made more precise by Newton.

The fundamental theorem related to order is due to Bezout. It states that *the number of points of intersection of a curve of order m and a curve of order n is mn*. Of course, here we must take into consideration multiplicities, complex points, and points at infinity.

Notwithstanding its importance, the classification of curves by order alone is rather crude for purposes of diophantine analysis. Two curves of the same order can have very very different sets of rational points. Thus the curve Γ with equation $x^2 + y^2 = 1$ has infinitely many rational points (with coordinates $x = 2k/(k^2 + 1)$, $y = (k^2 - 1)/(k^2 + 1)$, k rational), whereas the curve $x^2 + y^2 = 3$ has none.

The notion of greatest importance for diophantine analysis is that of birational equivalence of curves.

Definition 1. Two curves $f(x, y)$ and $g(u, v)$ are said to be *birationally equivalent* if the coordinates of each of them are expressible in terms of the coordinates of the other as rational functions with rational coefficients:

$$x = \varphi(u, v), \qquad u = \varphi_1(x, y),$$
$$y = \psi(u, v), \qquad v = \psi_1(x, y).$$

It is clear that the respective sets of rational points of two birationally equivalent curves coincide with the possible exception of a finite number of points. Birationally equivalent curves can have different orders, that is, the order of a curve is not a birational invariant. For example, the quartic curve

$$y^2 = x^4 - x^3 + 2x = (x - 1)(x^3 + 2)$$

can be transformed by means of the substitution

$$x = (1 + u)/u, \qquad y = v/u^2$$

into the cubic

$$v^2 = 3u^3 + 3u^2 + 3u + 1,$$

with u and v rationally expressible in terms of x and y :

$$u = 1/(x - 1), \qquad v = y/(x - 1)^2.$$

We will see that a quadratic curve with at least one rational point is birationally equivalent to a rational straight line.

It was Henri Poincaré who first called attention to the fundamental significance of birational transformations in the study of the arithmetic of algebraic curves. In the introduction to his famous paper "On the arithmetical properties of algebraic curves" he wrote: "I asked myself if it is not possible to connect many problems of

analysis on a systematic basis by introducing a new classification of homogeneous polynomials of higher order, analogous in a sense to the classification of quadratic forms.

This classification would have to be built on the foundation of the group of birational transformations admitted by the algebraic curve."[2]

One of the basic invariants of the group of birational transformations is the genus of a curve. To define it, we introduce first the notion of the simplest double point on a curve.

The *singular points* on a curve Γ given by (1) are the points whose coordinates satisfy the equations

$$f_x(x, y) = 0, \qquad f_y(x, y) = 0.$$

An algebraic curve has only finitely many such points. A singular point $P(x_0, y_0)$ is called a *double point* if at least one of the second partial derivatives f_{xx}, f_{xy}, and f_{yy} does not vanish at P. Finally, a *simplest double point* is a double point at which the curve has two noncoincident tangents (see Figure 1). When defining the genus of a curve we will assume that its only singular points are simplest double points. This is not a serious restriction, for it can be shown that an algebraic curve is birationally equivalent to one with only simplest double points.

We can now define the genus of a curve.

Definition 2. By the *genus* of a plane algebraic curve Γ of order n we mean the number

$$g = \frac{(n-1)(n-2)}{2} - d, \tag{3}$$

where d is the number of simplest double points on the curve.

It is clear that g is an integer. It can be shown that $g \geq 0$. If the order is 1 or 2, then $g = 0$. Such curves are called *rational.* The reason for this is that if a curve Γ of genus 0 with equation

$$F(x, y) = 0$$

has a rational point $P(x_0, y_0)$, then the coordinates x and y can be expressed in the form $x = \varphi(t)$, $y = \psi(t)$, where φ and ψ are rational functions with rational coefficients and $F(\varphi(t), \psi(t)) \equiv 0$. Moreover, $t = \chi(x, y)$, where χ is also a rational function with rational coefficients.

One also says that curves of genus 0 can be *uniformized by means of rational functions*.

If $n = 1$, that is, in the case of straight lines, it is clear that any two rational straight lines $Ax + By + C = 0$ and $A_1x + B_1y + C_1 = 0$ are birationally equivalent, that is, there is just one class of birationally equivalent straight lines.

If $n = 2$, that is, if the curve is a conic section, and if there is a rational point $P(x_0, y_0)$ on it, then the curve is *birationally equivalent to a straight line*. To see this it suffices to take an arbitrary rational straight line D and to establish a one-to-one correspondence between the points M on the conic and the points M' on D so that the points P, M, M' are collinear. Since every conic with a rational point is equivalent to a rational line, all such conics are (birationally) equivalent to each other, and so form a single class that includes all rational lines. This implies that if a conic has a rational point then it has infinitely many rational points.

There are a great many equivalence classes of conics without rational points.

Poincaré proved the following theorem: "Every curve of genus 0 and order $n > 2$ is birationally equivalent to a curve of order $n - 2$." (Hilbert and Hurwitz

proved a similar result 10 years earlier; see [3].) Hence a rational curve of genus 0 is always equivalent to a straight line or to a conic.

If a cubic curve has genus 0, then

$$\frac{(3-1)(3-2)}{2} - d = 0,$$

that is, $d = 1$. But then the curve must have a simplest double point which is clearly rational. A straight line passing through the double point P intersects the curve Γ in just one other point. We will show that in this case the cubic Γ is birationally equivalent to a rational line. To this end we take a rational line D and establish a one-to-one correspondence between the points M on Γ and the points M' on D so that the three points M, M', and the double point P are collinear (see Figure 2).

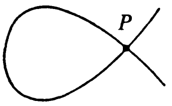


Figure 1

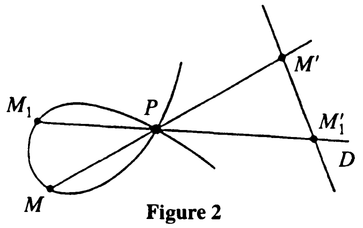


Figure 2

This shows that a cubic curve Γ with a simplest double point is birationally equivalent to a rational line. As such, it can be uniformized by means of rational functions. For example, $P(0,0)$ is a double point on the curve $y^2 = x^3 - 2x^2$. If we pass through P the lines $y = kx$, then we obtain $k^2x^2 = x^3 - 2x^2$, whence $x = k^2 + 2$, $y = k(k^2 + 2)$.

Now we consider curves of genus 1.

It can be shown that curves of genus 1 cannot be uniformized by means of rational functions but can be represented by means of elliptic functions of one variable. Hence the name *elliptic* is attached to such curves.

If an equation

$$f(x, y) = 0 \tag{4}$$

determines a curve Γ of genus 1 with a rational point $P(x_0, y_0)$, then it is possible to reduce it by means of birational transformations to the form

$$y^2 = x^3 + ax + b. \tag{5}$$

This is the so-called *Weierstrass normal form*. In this case it is possible to express x and y in terms of the Weierstrass functions

$$x = \wp(t), \qquad y = \wp'(t).$$

Thus the coordinates of the rational points on a cubic curve cannot, in general, be expressed as rational functions of a single parameter. However, if we know one or two rational points on such a curve, then we can find yet another rational point on it. To do this one makes use of two methods, known, respectively, as the *method of tangents* and the *method of secants*.

1. If P is a rational point on a cubic Γ, then by drawing at P the tangent to Γ we obtain a rational straight line (the slope of this line is rational) that intersects Γ in a third rational point. This is the *method of tangents*.

2. If P_1 and P_2 are rational points on Γ, then the rational line P_1P_2 intersects Γ in a third rational point P_3. This is the *method of secants*.

The fundamental theorem about curves of genus 1 was proved by Poincaré. It asserts that: **Every rational curve of genus 1 with a rational point is birationally equivalent to a cubic curve.**

Thus cubics are a model for the study of the arithmetic of curves of genus 1.

Let $\mathscr{M}$ be the set of rational points on an elliptic curve. Using the tangent and secant methods it is possible to impose on it the structure of an abelian group. In essence, this was done by Jacobi in 1835 in **[4]**. A deeper study of this group was carried out by Poincaré **[2]**, who surmised that this group has a finite number of generators. He called this number the *rank* of the cubic curve. It was later shown that the rank of a curve is an invariant of the group of birational transformations. Poincaré posed the question of the possible values of the rank of a cubic curve. This question remains open. The English mathematician L. J. Mordell proved the deep result that the rank of an elliptic curve is always finite.

Poincaré showed that the group of an elliptic curve can contain elements of finite order (that is, that it is a group with torsion). In essence, this was already known to Fermat and Euler.

We conclude by considering the geometric sense of the group operations associated with the method of secants and the method of tangents. We first reduce a cubic curve Γ with rational points to the form (5). Let A and B be rational points on Γ and let C' be the point in which the straight line AB intersects Γ. Then we call the point C, symmetric to C' with respect to the x-axis, the sum of the points A and B:

$$A \oplus B = C.$$

Thus if C' has coordinates (x, y), then C will have coordinates $(x\ -y)$. The transition from C' to C is of vital importance. It is only then that the operation of addition acquires the group properties, namely associativity, the existence of a zero element, and the existence of an additive inverse for each of its elements. The commutativity property of our operation is obvious.

To add A to itself, that is, to obtain $2A$, we use the method of tangents. We define the point $2A = D$ to be the point symmetric to the point D' in which the tangent at A intersects Γ.

It remains to find the point that plays the role of the zero element. No finite point will do. When we go to homogeneous coordinates by putting $x = u/w$, $y = v/w$, (5) becomes

$$v^2w = u^3 + auw^2 + bw^3. \tag{6}$$

If w and u are 0, then v is arbitrary and we can put $v = 1$. We denote by $\mathscr{O}$ the point at infinity on Γ with coordinates $(0, 1, 0)$. It is clear that the point $\mathscr{O}'$, symmetric to the point $\mathscr{O}$ with respect to the axis of abscissas, coincides with $\mathscr{O}$.

We show that $\mathscr{O}$ plays the role of zero. To this end we show that all vertical lines $u = mw$ intersect at $\mathscr{O}$. Indeed, if w and u are 0, then we can put $v = 1$. Now let $A(x_0, y_0)$ be a rational point on Γ. Then, according to what has just been shown, the straight line through A and $\mathscr{O}$ is vertical, that is, its equation is $x = x_0$. This straight line intersects Γ in the three points A, $\mathscr{O}$, and $A'(x_0, -y_0)$, the latter symmetric to A with respect to the x-axis. According to our definition, the sum of the points A and $\mathscr{O}$ is the point symmetric to A', that is, A itself. Thus $A \oplus \mathscr{O} = A$.

Finally, the inverse of A is $A'(x_0, -y_0)$. Indeed, the straight line joining these two points is vertical, and so intersects Γ at $\mathcal{O}$, that is, $A \oplus A' = \mathcal{O}$.

Points of finite order are characterized by the fact that $nA \Leftarrow A$ for some n, that is, there is a return to the initial point.

REFERENCES

1. F. Kirwan, *Complex algebraic curves*, London Math. Soc. Student Texts 23, Cambridge (1992). (Included by the translator.)
2. H. Poincaré, Sur les propriétés arithmetiques des courbes algèbriques, *J. Math.*, 5ᵉ série, 1901, **7**, 161–233.
3. D. Hilbert, A. Hurwitz, Über die diophantische Gleichungen vom Geschlecht Null, *Acta Math.*, 1890, **14**, 217–224.
4. C. G. J. Jacobi, De usu theoriae integralium ellipticorum et integralium abelia-norum in analysi Diophantea, *Crelle's Journal für die reine und angew. Mathematik*, 1835, **13**, 353–355. Gesammelte Werke, Bd. 2, S.53–55.
5. J. H. Silverman, J. Tate, *Rational points on elliptic curves*, Springer (1992). (Included by the translator.)

Albert Nijenhuis and Paul Erdős (1913–1996)
Combinatorics and Graph Theory Conference
in honor of Herb Wilf's 65th birthday
June 12–15, 1996, University of Pennsylvania
Photo by Stan Wagon, Macalester College

PROBLEMS AND SOLUTIONS

Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**

with the collaboration of Paul T. Bateman, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Underwood Dudley, Michael A. Filaseta, Ira M. Gessel, Bart Goddard, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Robert Israel, Murray S. Klamkin, Daniel J. Kleitman, Fred Kochman, Frederick W. Luttmann, Frank B. Miles, M. J. Pelling, Richard Pfiefer, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the* MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before June 30, 1997; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.

PROBLEMS

10564. *Proposed by Aviezri Fraenkel, Weizmann Institute of Science, Rehovot, Israel.* The *Nim-sum* of two positive integers with binary expansions $\sum_{i\geq 0} a_i 2^i$ and $\sum_{i\geq 0} b_i 2^i$ is the number with binary expansion $\sum_{i\geq 0} c_i 2^i$, where a_i, b_i, c_i are in $\{0, 1\}$ and $c_i \equiv a_i + b_i \bmod 2$. Let n be a positive integer and let j be a nonnegative integer. How many of the 2^n subsets of the set $\{1, 2, \ldots, n\}$ have the property that their elements have Nim-sum equal to j?

10565. *Proposed by D. M. Bloom, Brooklyn College, Brooklyn, New York and Kenneth Suman, Winona State University, Winona, MN.* A rectangle is composed of mn squares arranged in m rows and n columns. In a certain game, the squares are selected one by one at random (without replacement). What is the expected number of selections until j columns of the rectangle are composed entirely of selected squares? (When $j = 1$, $m = 5$, and $n = 15$, the problem asks for the expected length of a type of bingo game known as a line game.)

10566. *Proposed by Gerry Myerson, Macquarie University, Australia.* Let S be a finite set of cardinality $n > 1$. Let f be a real-valued function on the power set of S, and suppose $f(A \cap B) = \min\{f(A), f(B)\}$ for all subsets A and B of S. Prove that

$$\sum (-1)^{n-|A|} f(A) = f(S) - \max f(A),$$

where the sum is taken over all subsets A of S and the maximum is taken over all proper subsets A of S.

10567. *Proposed by Donald Girod, Canisius College, Buffalo, NY.* Let $f : [0, 1] \to \mathbb{R}$ be a continuous function with $f(0) = f(1) = 0$. Show that the Lebesgue measure of $\{h : f(x + h) = f(x) \text{ for some } x \in [0, 1]\}$ is at least $1/2$.

10568. *Proposed by Donald E. Knuth, Stanford University, Stanford, CA.* Let n be a nonnegative integer. The sequence defined by $x_0 = n$ and $x_{k+1} = x_k - \lceil \sqrt{x_k} \rceil$ for $k \geq 0$ converges to 0. Let $f(n)$ be the number of steps required; i.e., $x_{f(n)} = 0$ but $x_{f(n)-1} > 0$. Find a closed form for $f(n)$.

10569. *Proposed by W. M. Priestley, University of the South, Sewanee, TN.* Let X and Y be countable subsets of real numbers (each endowed with the subspace topology). If there exist one-to-one continuous maps of X onto Y and of Y onto X, does it follow that X and Y are homeomorphic?

10570. *Proposed by Emeric Deutsch, Polytechnic University, Brooklyn, NY.* An ordered tree is a rooted tree in which the children of each node form a sequence rather than a set. The height of an ordered tree is the number of edges on a path of maximum length starting at the root. Let $a(n, k)$ denote the number of ordered trees with n edges and height k and let $S(n, k)$ be the Stirling numbers of the second kind (the number of partitions of $\{1, 2, \ldots, n\}$ into k nonempty parts). Note that $a(n, 1) = S(n, 1)$, since both numbers are 1. Show that **(a)** $a(n, 2) = S(n, 2)$, **(b)** $a(n, 3) + a(n, 4) = S(n, 3)$, and **(c)*** generalize these observations.

SOLUTIONS

An Arithmetic Function of Modest Size

10192 [1992, 61]. *Proposed by the late Paul Erdős, Hungarian Academy of Sciences, Budapest, Hungary.* Let $L(n)$ denote the least common multiple of the positive integers not exceeding n. For $n \geq 2$ let $g(n)$ denote the largest positive integer k such that $n^k \mid L(n)$. For example, $g(2) = 1$, $g(30) = 2$, $g(420) = 3$. Prove that for x large

$$\max_{2 \leq n \leq x} g(n) = \log x / \{\log\log x + o(1)\}.$$

Solution by the editors based on the solutions of Richard Stong, Rice University, Houston, TX, and the proposer. We use the prime number theorem in the following two forms

$$\lim_{t \to \infty} \frac{\pi(t)}{t/\log t} = 1, \quad \lim_{t \to \infty} \frac{\theta(t)}{t} = 1,$$

as well as Chebyshev's theorem that $\pi(2j) > \pi(j)$ for $j = 1, 2, 3, \ldots$. Here $\pi(t)$ is the number of prime numbers not exceeding t and $\theta(t)$ is the sum of the logarithms of the primes not exceeding t.

Clearly

$$L(n) = \prod_p p^{\lfloor \log n / \log p \rfloor},$$

where p runs through the prime numbers not exceeding n. Thus if $p^\alpha \mid n$ but $p^{\alpha+1} \nmid n$ for some prime p, and if $n^g \mid L(n)$, then $g\alpha \leq \log n / \log p$ or $g \leq \log n / \log p^\alpha$. Hence if q_n is the largest prime power dividing n, we have $g(n) = \lfloor \log n / \log q_n \rfloor \leq \omega(n)$, where $\omega(n)$ is the number of distinct prime factors of n. (Actually $g(n) < \omega(n)$ whenever $\omega(n) > 1$.)

Now suppose we are given a positive number ϵ. If $q_n \geq e^{-\epsilon} \log n$, then

$$g(n) = \left\lfloor \frac{\log n}{\log q_n} \right\rfloor \leq \frac{\log n}{\log(e^{-\epsilon} \log n)} = \frac{\log n}{\log\log n - \epsilon}.$$

If $q_n < e^{-\epsilon} \log n$, then $g(n) \leq \omega(n) \leq \pi(q_n) \leq \pi(e^{-\epsilon} \log n)$. Since $e^{-\epsilon} < 1$, the prime number theorem gives $\pi(e^{-\epsilon} \log n) < \log n / \log\log n$ if n is sufficiently large. Thus in either case $g(n) \leq \log n / (\log\log n - \epsilon)$, provided n is sufficiently large. Since $\log x / (\log\log x - \epsilon)$ is an increasing function of x for $x > \exp\exp(1 + \epsilon)$, we obtain

$$\max_{2 \leq n \leq x} g(n) \leq \log x / (\log\log x - \epsilon). \tag{1}$$

for large x.

To get a lower bound let $p_1, p_2, \ldots$ be the primes in increasing order. If x is a positive number greater than 2, let k be the largest integer such that $p_1 p_2 \cdots p_k \le x$, i.e., such that $\theta(p_k) \le \log x$. If $m = p_1 p_2 \cdots p_k$, then $g(m) = \lfloor \log m / \log p_k \rfloor$. Since $x < p_{k+1} m < 2p_k m \le p_k^2 m$, we have $\log x < \log m + 2\log p_k$, so that $g(m) \ge (\log m / \log p_k) - 1 > (\log x / \log p_k) - 3$. If x is large, then k is large and so the prime number theorem gives $\log x \ge \theta(p_k) > e^{-\epsilon/2} p_k$ or $\log p_k < \log\log x + \epsilon/2$. Hence

$$\begin{aligned}\max_{2\le n\le x} g(n) &\ge g(m)\\ &> \log x \Big/ \big(\log\log x + \epsilon/2\big) - 3\\ &> \log x \Big/ \big(\log\log x + \epsilon\big),\end{aligned} \tag{2}$$

provided x is sufficiently large. In view of (1) and (2), the assertion of the problem is established.

Although $g(n)$ is sometimes much smaller than $\omega(n)$, a similar argument gives

$$\max_{2\le n\le x} \omega(n) = \frac{\log x}{\log\log x - 1 + o(1)}.$$

While the value of n in the interval $[2, x]$ for which $\omega(n)$ is largest is obvious, such is not the case for $g(n)$.

Paul Erdős reminded the editors that in his original proposal he put forward only the weaker assertion

$$\max_{2\le n\le x} g(n) = (1 + o(1)) \log x \Big/ \log\log x,$$

which is a little easier to prove than the assertion of the problem as published. Erdős also remarked that it would be of interest to determine the sequence of "champions" for the arithmetical function g, i.e., to determine for each k the smallest integer n_k greater than 1 for which $g(n_k) = k$. For example, $n_1 = 2$, $n_2 = 30$, $n_3 = 420$, $n_4 = 27720$. Since $g(n) < \omega(n)$ when n is not a prime power, n_k must have at least $k + 1$ prime factors when $k > 1$.

Solved also by L. E. Mattics.

World Series, 1994

10223 [1992, 462]. *Proposed by Julio Kuplinsky, Amherst, NY.* For $p \in \mathbb{R}$, $q = 1 - p$, and positive integers n, prove

$$\sum_{k=n}^{2n-1} \binom{k-1}{n-1}\Big[p^n q^{k-n} + p^{k-n} q^n\Big] = 1.$$

Editorial comment. The challenge here was to give a proof that is valid for all $p \in \mathbb{R}$, since problem E3386 [1990, 427; 1992, 22] obtained this result for $0 \le p \le 1$ through a probabilistic interpretation. Several readers noted that the sum on the left side of the desired identity defines a polynomial in p, so the general identity follows from its truth on any infinite set. Proofs avoiding the probabilistic interpretation can be constructed by rewriting $\binom{k-1}{n-1}$ as $\binom{k}{n} - \binom{k-1}{n-1}$ and rearranging terms to allow the sum for n to be compared to that for $n + 1$.

Frank Schmidt has noted that the result also follows from the additional remarks to the solution of E1829 [1965, 1201; 1967, 323; 1967, 1134], establishing the identity

$$\sum_{k=n}^{r} \binom{k-1}{n-1} x^n (1-x)^{k-n} = \sum_{k=n}^{r} \binom{r}{k} x^k (1-x)^{r-k}$$

for all $r \geq n$. Problem E2681 [1977, 728; 1979, 129] is also related. Murray S. Klamkin pointed out that a general two-variable form appears in *Crux Mathematicorum*, Problem 183 [1976, 193; 1977, 69], and a form with an arbitrary number of variables appears in *SIAM Review*, Problem 85-10, 28 (1986), p. 243. Furthermore, a solution to a variant of this problem has already appeared in these pages: see Doron Zeilberger, On an identity of Daubechies, this MONTHLY 100 (1993), 487. Earlier appearances of the problem include solutions of the two types outlined.

Solved by J. Anglesio (France), D. Beckwith, J. C. Binz (Switzerland), M. Bowron, D. Callan, R. J. Chapman (U. K.), W. Chu (China), P. Deiermann, S. B. Ekhad, J. Fukuta (Japan), D. A. Grable, C. P. Grant, P. Griffin, H. van Haeringen (The Netherlands), R. Holzsager, W. K. Jeong (Korea), A. M. Karparvar (Iran), M. S. Klamkin (Canada), B. G. Klein, N. Komanda, D. W. Koster, I. I. Kotlarski, R. A. Leslie, O. P. Lossers (The Netherlands), M. Mócsy (Hungary), I. Nemes (Austria), K. Perera, C. R. Pranesachar (India), P. Ranaldi, R. W. Richards, R. Richberg (Germany), J. B. Robertson, F. Schmidt, J. H. Steelman, H. L. Stubbs, L. Verde-Star (Mexico), M. Vowe (Switzerland), H. S. Wilf, Centre Problem Solving Group, and the proposer.

Primitive Elements Modulo Primes and Their Squares

10311 [1993, 499]. *Proposed by Solomon W. Golomb, University of Southern California, Los Angeles, CA.* It is well-known that if g is a primitive root modulo p, where $p > 2$ is prime, either g or $g + p$ (or both) is a primitive root modulo p^2 (indeed modulo p^k for all $k \geq 1$.)
(**a**) Find an example of a prime $p > 2$ and a primitive root g modulo p with $1 < g < p$ such that g is *not* a primitive root modulo p^2.
(**b**) Show that, among all $\phi(p-1)$ primitive roots g modulo p with $1 < g < p$, at least half of them are also primitive roots modulo p^2.

Editorial comment. For part (**a**), the most popular solution was $(p, g) = (29, 14)$ for which p is minimal. The example $(p, g) = (40487, 5)$ (see E. L. Litver & G. E. Judina, Primitive roots for the first million primes and their powers, *Mathematical Analysis and Its Applications* 3 (1971), 106–109, Izdat. Rostov. Univ.) was given as the only example with p less than one million in which g is the smallest primitive root modulo p. Other examples given by readers were $(p, g) = (71, 11)$, $(37, 18)$ and $(487, 10)$. The latter example appears in several books (e.g., D. Shanks, *Solved and Unsolved Problems in Number Theory*, Chelsea, 1985, ex. 79, p. 102).

A weaker version of (**b**) appeared in *Math. Magazine* as Problem 1419 [1993, 126; 1994, 148]. The solution published there could be modified to solve this problem. That argument is similar to one found in Alfred Brauer, Elementary estimates for the least primitive root, *Studies in mathematics and mechanics presented to Richard von Mises*, Academic Press, 1954, 20–29, where the result appears as Theorem 7. From there, the method can be traced back to V. A. Lebesgue, Théorème sur les racines primitives, *Comptes Rendus Acad. Sci. Paris* 64 (1867), 1268–1269. Following a suggestion of Paul Bateman, we reprint Lebesgue's exact words.

Soit $g < p$ une racine primitive pour le module premier p; soit encore $g' < p$ et $g' \equiv g^{p-2}$ (mod. p). Le nombre $g' < p$ sera aussi racine primitive. Ces racines g, g', satisfaisant à la condition $gg' \equiv 1$ (mod. p), sont associées. L'une d'elles au moins est racine primitive pour le module p^n, quel que soit l'exposant n.

No proof is given there, but this statement was introduced with the words: "La démonstation ne présente pas de difficultés."

Published tables are useful in finding the examples requested in (**a**). In particular, for $g < 100$ and $p < 2^{32}$ one need identify only the primitive roots among the values given in Peter L. Montgomery, New solutions of $a^{p-1} \equiv 1 \pmod{p^2}$, *Math. Comp.* 61 (1993), 361–363. This is an easy exercise for a computer algebra system. Thus, the pair $(p, g) = (113, 68)$ is seen to be an example of a primitive root g modulo p and $g^{p-1} \equiv 1 \pmod{p^3}$.

This problem suggests study of the set $G(p)$ of numbers g that are primitive roots modulo p with $0 < g < p$, but fail to be primitive roots modulo p^2. Some readers included the results of computational work determining $G(p)$ for all p in certain intervals. In particular, Albert Wassermann included a complete table of the sets $G(p)$ with $p < 1000$, and John P. Robertson summarized a computer search of $2 < p < 20000$. This range was divided into eight subintervals and the number of p in each subinterval with each possible size of $G(p)$ was given. Some noteworthy examples of $G(p)$ were also included. For example, $G(653) = \{84, 120, 287, 410\}$ and $G(16631) = \{274, 11047, 14697, 16026\}$ are the only examples of $\#\big(G(p)\big) \geq 4$ in this range.

Numerical evidence suggests that $\#\big(G(p)\big)$ is really much smaller than $\phi(p-1)/2$. Such a result is known. Paul Bateman also provided a reference to S. D. Cohen, R. W. K. Odoni, and W. W. Strothers, On the least primitive root modulo p^2, *Bull. London Math. Soc.* 6 (1974), 42–46, where it is shown that, for any $c > 1/2$, there is a quantity $P(c)$ such that $p > P(c)$ implies that $\#\big(G(p)\big) < p^c$.

Solved by D. Alvis, R. Barbara (Lebanon), P. T. Bateman & W. P. Wardlaw, K. A. Beres, V. Božin (Yugoslavia), D. Callan (part b only), R. J. Chapman (U. K.), J. Christopher, H. M. Edgar , H. S. Gunaratne (Brunei), W. Johnson, I. Kastanas, D. W. Koster, Y. H. Kwong, D. E. Manes, M. Newman (Israel, part b only), J. P. Robertson, R. M. Robinson, H. Schmidt Jr. (part a only), R. Simion & F. Schmidt, J. H. Steelman, A. Wassermann (Germany, part a only), GCHQ Problem Solving Group (U. K.), the MMRS group of Oklahoma State University, and the proposer.

Asymptotics in Three Parts

10335 [1993, 797]. *Proposed by David Borwein, University of Western Ontario, London, Ontario, Canada, and Jonathan Borwein, Simon Fraser University, Burnaby, British Columbia, Canada.* Let r be a positive constant and $c_0 \geq 0$. Consider the iteration $c_{n+1} = c_n + r - c_n/\sqrt{1+c_n^2}$. **(a)** For which values of r does the sequence $\langle c_n \rangle$ converge? **(b)** In case of convergence to c with $c \neq c_0$, prove that $\lim(c_{n+1} - c)/(c_n - c)$ exists and determine its value. **(c)** In case of divergence, find an asymptotic expression for c_n.

Solution for $r \neq 1$ by Heinz-Jürgen Seiffert, Berlin, Germany. More generally, for $r > 0$, $k > 0$, and $c_0 \geq 0$, we consider the iteration

$$c_{n+1} = c_n + r - \frac{c_n}{\left(1 + c_n^k\right)^{1/k}}.$$

(a) If $\langle c_n \rangle$ converges to c, then $r = c/\left(1+c^k\right)^{1/k} < 1$. Hence, the condition $0 < r < 1$ is necessary for the convergence of $\langle c_n \rangle$. We show that it is also sufficient. Let $0 < r < 1$ and define $M = \max\left\{c_0, r + r/\left(1-r^k\right)^{1/k}\right\}$. Consider the mapping $T : [0, M] \to \mathbb{R}$ defined by $T(x) = x + r - x/\left(1+x^k\right)^{1/k}$. If $0 \leq x \leq r/\left(1-r^k\right)^{1/k}$, then $0 < r \leq T(x) \leq x + r \leq M$. If $r/\left(1-r^k\right)^{1/k} \leq x \leq M$, then $0 < r \leq T(x) \leq x \leq M$. Thus, we have $T\big([0, M]\big) \subseteq [0, M]$. Furthermore, for all $x \in [0, M]$,

$$0 \leq T'(x) = 1 - \frac{1}{\left(1+x^k\right)^{1+1/k}} \leq 1 - \frac{1}{\left(1+M^k\right)^{1+1/k}},$$

so T is a contraction. Since $c_0 \in [0, M]$ and $c_{n+1} = T(c_n)$, the contraction mapping principle implies that $\langle c_n \rangle$ converges to the unique fixed point $c \in [0, M]$ of T. The equation $T(c) = c$ is easily solved, giving $c = r/\left(1-r^k\right)^{1/k}$.

(b) Let $0 < r < 1$ and $c = r/\left(1-r^k\right)^{1/k} \neq c_0$. Since $T'(x) > 0$, all $c_n \neq c$ and

$$\begin{aligned}\lim_{n\to\infty} \frac{c_{n+1}-c}{c_n - c} &= \lim_{n\to\infty} \frac{T(c_n) - T(c)}{c_n - c} = T'(c) \\ &= 1 - \frac{1}{\left(1+c^k\right)^{1+1/k}} = 1 - \left(1-r^k\right)^{1+1/k}.\end{aligned}$$

(c) Let $r > 1$ and $k > 0$. Using the obvious inequality $c_{n+1} \geq c_n + r - 1$, for $n \geq 0$, an easy induction gives $c_n \geq (r-1)n$ for all $n \geq 0$. Now, $1 - c_n/\left(1 + c_n^k\right)^{1/k} = 1 - \left(1 + c_n^{-k}\right)^{-1/k} = O\left(c_n^{-k}\right)$, for $n \geq 1$, and it now follows that $c_{n+1} = c_n + r - 1 + O(n^{-k})$. Hence, for all $n \geq 2$,

$$c_n = \begin{cases} (r-1)n + O(n^{1-k}) & \text{if } 0 < k < 1, \\ (r-1)n + O(\log n) & \text{if } k = 1, \\ (r-1)n + O(1) & \text{if } k > 1. \end{cases}$$

Thus for $r > 1$ and $k > 0$, we have $c_n \sim (r-1)n$ as $n \to \infty$.

When $r = 1$, we conjecture that for all $k > 0$, $c_n \sim ((k+1)n/k)^{1/(k+1)}$ as $n \to \infty$.

Solution for $r = 1$ (and $k = 2$) by Robert D. Brown & Pawel Szeptycki, University of Kansas, Lawrence, KS. We show that $c_n n^{-1/3} \to (1.5)^{1/3}$ as $n \to \infty$ (this can be guessed by considering the differential equation $x' = 0.5x^{-2}$ suggested by the following estimates).

We have

$$c_{n+1} = c_n + \frac{1}{\sqrt{1 + c_n^2}\left(\sqrt{1 + c_n^2} + c_n\right)}$$

and so $L(c_n) < c_{n+1} < R(c_n)$ with $L(x) = x + 1/(2 + 2x^2)$ and $R(x) = x + 1/(2x^2)$. Both $L(x)$ and $R(x)$ are increasing for $x \geq 1$. Let $\epsilon > 0$. Choose n_ϵ so that

$$\frac{1 - 2\epsilon/3}{x^{2/3}} \leq \frac{1}{1 + x^{2/3}}$$

for $x \geq (1.5 - \epsilon)n_\epsilon$, and then choose m_ϵ so that $c_{m_\epsilon} > \sqrt[3]{(1.5 - \epsilon)n_\epsilon}$. Use induction on n to show that

$$c_{n+m_\epsilon} > \sqrt[3]{(1.5 - \epsilon)(n + n_\epsilon)} \tag{$*$}$$

for all $n \geq 0$; the case $n = 0$ follows from our definitions. Suppose $(*)$ holds for n, and let $x = (1.5 - \epsilon)(n + n_\epsilon)$. The case $n + 1$ of $(*)$ holds if $\sqrt[3]{x + 1.5 - \epsilon} < L(\sqrt[3]{x})$. However, since $\sqrt[3]{a + b} - \sqrt[3]{a} < b/\left(3\sqrt[3]{a^2}\right)$ for $a, b > 0$, this follows from the choice of n_ϵ. Thus $\liminf_{n\to\infty} c_n/\sqrt[3]{n} \geq \sqrt[3]{1.5 - \epsilon}$. Similarly, choose k_ϵ so that $c_0 < \sqrt[3]{(1.5 + \epsilon)k_\epsilon}$ and $(1 + 2\epsilon/3)/(x + 1.5 + \epsilon)^{2/3} > x^{-2/3}$ for all $x \geq (1.5 + \epsilon)k_\epsilon$. A similar induction shows that $c_n < \sqrt[3]{(1.5 + \epsilon)(n + k_\epsilon)}$. Thus $\limsup_{n\to\infty} c_n/\sqrt[3]{n} \leq \sqrt[3]{1.5 + \epsilon}$.

Editorial comment. Kiran Kedlaya used an asymptotic expansion of $T(x)$ to determine a differential equation that could be used to guess the asymptotic behavior of c_n for $r = 1$. Allan Pedersen formulated

Lemma. *Let a sequence $\langle x_n \rangle$ be defined by $x_0 = c_0 \geq 0$ and $x_{n+1} = x_n + g(x_n)$ for $n \geq 0$, where $g(x)$ is a continuous, positive, decreasing function for $x \geq 0$. Let $x(t)$ be the solution to the initial value problem*

$$x(0) = c_0, \ \frac{dx}{dt} = g(x), \ t \geq 0.$$

Then $x_n \to \infty$ and $x_n = x(n) + O(1)$ as $n \to \infty$.

Proof. See MONTHLY Problem 6610 [1989, 744; 1991, 448].

Although an explicit solution of the differential equation is possible in this case, its main use is to obtain an asymptotic expression. O. P. Lossers remarked that for $r > 1$, one can prove the validity of an expansion

$$c_n = n(r-1) + d + \sum_{k=1}^{\infty} a_k n^{-k}$$

as $n \to \infty$, whose coefficients could be calculated by substitution in the recurrence. A similar process for $r = 1$ gives

$$c_n^3 = \frac{3}{2}n - \frac{27}{8}n^{1/3} + \frac{8}{9}\log n + K + O(n^{-1/3})$$

where K is a constant.

Solved also by J. Anglesio (France), P. Bracken (Canada), R. J. Chapman (U. K.), D. A. Darling, D. Doster, M. Dresević (Yugoslavia), N. Eklund, J. Ferrer (Spain), R. A. Groeneveld, H. S. Gunaratne (Brunei), R. Holzsager, K. S. Kedlaya, P. G. Kirmser, E. H. Larson, K.-W. Lau (Hong Kong), O. P. Lossers (The Netherlands), B. Margolis (France), C. A. Minh, A. Pedersen (Denmark), I. A. Sakmar (Turkey), K. Schilling, M. Vowe (Switzerland), Z. Zhang (China), GCHQ Problem Solving Group (U. K.), NSA Problems Group, and the proposer.

Expected Number of Sums in a Given Set

10336 [1993, 797]. *Proposed by Ignacy I. Kotlarski, Oklahoma State University, Stillwater, OK.* Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables, each exponentially distributed with parameter a, $a > 0$, i. e., for $k = 1, 2, \ldots$,

$$Pr(X_k \le x) = \begin{cases} 0 & \text{if } x \le 0, \\ 1 - e^{-ax} & \text{if } x > 0. \end{cases}$$

Let B be a fixed Borel set in $[0, \infty)$ such that its Lebesgue measure $\mu_L(B)$ is finite and positive. Let $Y_k = X_1 + \cdots + X_k$ for $k = 1, 2, \ldots$, and $\theta = \sum_{k=1}^{\infty} Pr(Y_k \in B)$.
(**a**) Find θ as a function of a.
(**b**) Find a uniform minimum variance unbiased estimator of θ from a sample from the above exponential distribution of a fixed size n.

Solution I of (**a**) *by Robert A. Agnew, FMC Corporation, Chicago, IL.* $\theta = a\mu_L(B)$. It is well known that the probability density function of Y_k on $[0, \infty)$ is $f_k(y) = a^k y^{k-1} e^{-ay}/(k-1)!$ (the gamma distribution). Hence $Pr(Y_k \in B) = \int_B f_k(y)\,dy$ and

$$\theta = \sum_{k=1}^{\infty} \int_B f_k(y)\,dy = \int_B \sum_{k=1}^{\infty} f_k(y)\,dy = \int_B a\,dy = a\mu_L(B).$$

Solution II of (**a**) *by Kenneth Schilling, University of Michigan, Flint, MI.* We prove that $\theta = a\mu_L(B)$. The function $\theta(B) = \sum_{k=1}^{\infty} Pr(Y_k \in B)$ is a countably additive measure on the Borel sets, so it suffices to prove the claim when B is the interval $[0, t)$ for $t > 0$.

Let $\{Z_t\}_{t \ge 0}$ be the Poisson process whose interarrival times are $\{X_1, X_2, \ldots\}$, that is, $Z_t = \max\{k : Y_k < t\}$. For $t > 0$, the random variable Z_t has a Poisson distribution with mean at. Thus, if $B = [0, t)$, we have

$$\theta(B) = \sum_{k=1}^{\infty} Pr(Y_k < t) = \mathbf{E}\left(\sum_{k=1}^{\infty} 1_{\{Y_k < t\}}\right) = \mathbf{E}(Z_t) = at.$$

Solution of (**b**) *by Markus Roters, University of Trier, Trier, Germany.* For each $n \in \mathbb{N}$, the random variable Y_n has an Erlang distribution with parameters n and a, and is a complete and sufficient statistic for the family of joint distributions of $X_1, \ldots, X_n$ indexed by the parameter a. Hence the Rao-Blackwell and Lehmann-Scheffé theorems imply that if an unbiased estimator for θ exists, there also exists one depending on Y_n, which is then a uniform minimum variance unbiased estimator (UMVUE).

For $n = 1$, there is no unbiased estimator for θ. Indeed, if $d(X_1) = d(Y_1)$ were such an unbiased estimator, it would follow for all $a > 0$ that

$$\mathbf{E}\left(d(Y_1)\right) = \theta \text{ if and only if } \int_{[0,\infty)} d(y)e^{-ay}\,d\mu_L(y) = \mu_L(B),$$

and hence, by differentiating with respect to a, for all $a > 0$,

$$\int_{[0,\infty)} \big(-yd(y)\big)e^{-ay}d\mu_L(y) = 0.$$

But now, by completeness, we could conclude $d(Y_1) = 0$ a.s., an impossibility.

However, for $n \geq 2$, one computes $d^*(Y_n) = (n-1)\mu_L(B)/Y_n$ is unbiased for θ and hence is the desired UMVUE. In fact, for $n = 2$, any unbiased estimator Z must have infinite variance. Otherwise, by Rao-Blackwell, there would exist an unbiased estimator of the form $f(Y_2)$, which has finite variance since $\mathrm{Var}(f(Y_2)) \leq \mathrm{Var}(Z) < \infty$. But Y_2 is complete, so there is at most one unbiased estimate θ that is a function of Y_2. Since $\mu_L(B)/Y_2$ is an unbiased estimate of θ, it follows that $f(Y_2) = \mu_L(B)/Y_2$ a.s., but this is impossible since $1/Y_2$ has infinite variance. Thus no unbiased finite variance estimator exists. $\mathrm{Var}\, d^*(Y_n) < \infty$ if and only if $n \geq 3$.

Solved also by D. Callan, D. A. Darling, E. Hertz, C. Peters, G. S. Rogers, E. A. Weinstein, GCHQ Problem Solving Group (U. K.), and the proposer.

Generalizing "Every Even Number Is The Sum of Two Odds"

10338 [1993, 873]. *Proposed by Charles Vanden Eynden, Illinois State University, Normal, IL.* Given an integer $n > 1$, determine the set of integers which can be written as a sum of two integers relatively prime to n.

Solution I by Kevin Ford, University of Texas, Austin, TX. When n is even, the desired set is all even integers; when n is odd, it is all integers.

When n is even, the summands must be odd, and the sum of two odd numbers is even. When m and n are not both even, we provide a realization of m. Let $p_1, \ldots, p_k$ be the primes dividing n. For each i, let b_i be a number not congruent to 0 or to m modulo p_i. By the Chinese Remainder Theorem, there is a number h such that $h \equiv b_i \pmod{p_i}$ for all i. It follows that both h and $m - h$ are relatively prime to n.

Solution II by Nasha Komanda, Central Michigan University, Mt. Pleasant, MI. With n fixed, for any integer k let $q(k)$ be the product of all primes that divide n but not k (take $q(k) = 1$ if there are no such primes). Observe that $k + q(k)$ and $k - q(k)$ are relatively prime to n. As in Solution I, we must realize m when m and n are not both even. If m and n are odd, we use $\big(m + q(m)\big)\ /\ 2$ and $\big(m - q(m)\big)\ /\ 2$. If m is even, we use $m/2 + q(m/2)$ and $m/2 - q(m/2)$.

Editorial comment. Most solvers used the Chinese Remainder Theorem. Sydney Bulman-Fleming observed that the answer would be different if m were required to be the sum of two positive integers. With this interpretation, there would be infinitely many counterexamples (e.g., $(m, n) = (4, 6)$ or $(m, n) = (7, 15)$) to the result shown here. Frank Schmidt noted that a solution can be obtained using problem 49 in W. Sierpiński, *250 Problems in Number Theory*, Elsevier, U.S. edition 1970, which reads: "Prove that for every positive integer m every even integer $2k$ can be represented as a difference of two positive integers relatively prime to m."

Solved also by P. J. Anderson (Canada), R. Barbara (Lebanon), P. Budney, S. Bulman-Fleming (Canada), G. Ehrlich, C. Lanski, J. H. Lindsey II, O. P. Lossers (The Netherlands), A. Pedersen (Denmark), R. M. Robinson, F. Schmidt, A. Tissier (France), A. N. 't Woord (The Netherlands), NSA Problems Group, and the proposer.

Disjoint Connections

10341 [1993, 874]. *Proposed by George Cain and Zhiging Lu, Georgia Institute of Technology, Atlanta, GA.* Let $\mathbf{D} = \{(x, y) : x^2 + y^2 \leq 1\}$ be the unit disk in the plane, and let $\{A_1, A_2, \ldots, A_n\}$ be a pairwise disjoint collection of finite subsets of the set $\mathbf{C} =$

$\{(x, y) : x^2 + y^2 = 1\}$. Prove that there is a pairwise disjoint collection $\{K_1, K_2, \ldots, K_n\}$ of connected subsets of **D** such that $A_i \subset K_i$ for each $i = 1, 2, \ldots, n$.

Solution I by Eugene Curtin, Southwest Texas State University, San Marcos, TX. We prove the result without assuming the sets A_j are finite, or even that there are finitely many such sets. For each $\theta \in [0, 2\pi)$ let

$$B_\theta = \left\{ \left(r\cos\left(\theta + \tfrac{1}{1-r}\right), r\sin\left(\theta + \tfrac{1}{1-r}\right) \right) : 0 < r < 1 \right\}.$$

The sets $\{ B_\theta : \theta \in [0, 2\pi) \}$ are pairwise disjoint, and each B_θ has the entire unit circle contained in its limit set. If $\{ A_v : v \in V \}$ is any family of pairwise disjoint nonempty subsets of the unit circle, then for each $v \in V$ pick an angle $\theta(v)$ such that $\big(\cos\theta(v), \sin\theta(v)\big) \in A_v$. Then let $K_v = B_{\theta(v)} \cup A_v$. Each K_v is connected since it is the union of a connected set with a subset of its limit points. Thus, $\{ K_v : v \in V \}$ is a family of connected pairwise disjoint subsets of the disk with $A_v \subset K_v$ for all $v \in V$.

Solution II by Frank Schmidt, Arlington, VA. Given sets A_i as in the original statement, we sketch a construction of sets K_i.
Step 1. Draw polygonal arcs connecting the points in each A_i, $1 \le i \le n$.
Step 2. Modify these arcs, if necessary. to achieve general position. That is, no more than two arcs should cross at any point.
Step 3. If an arc joining points in A_i meets an arc joining points in A_j with $j \neq i$, remove the crossing as in R. J. MacG. Dawson, "Paradoxical connections", this MONTHLY 96 (1989), 31–33.

Editorial comment. The most common solution was a spiral construction like Solution I. Other solutions used comb spaces or the graph of $\sin(1/x)$. Leroy F. Meyers noted the connection with MONTHLY Problems E1515 [1962, 312; 1963, 95] and 10328 [1993, 689; 96, Nov].

Solved also by J. W. Grossman, R. Holzsager, S. N. Kass, U. Klein (Germany), L. F. Meyers, A. Mülller (Switzerland), S. Ott, G. Poor & R. Griffus, T. Richmond & B. Richmond, A. Riese, H. Schlais, A. W. Schurle, W. R. Smythe, S. T. Stefanov (Bulgaria), A. N. 't Woord (The Netherlands), the New Mexico Tech Problem Solving Group, NSA Problems Group, The Citadel Problem Solving Group, and the proposers.

Emergence of an Abelian Group

10342 [1993, 874]. *Proposed by Shmuel Rosset, Tel Aviv University, Ramat Aviv, Israel.* Let F be a free group, and let R be a normal subgroup of F. Consider the subgroups $[R, nF]$ defined by

$$[R, nF] = \begin{cases} R & \text{if } n = 0, \\ \big[\,[R, (n-1)F], F\,\big] & \text{if } n > 0. \end{cases}$$

Prove that the set of elements of finite order in $R/[R, nF]$ is an abelian group.

Solution by Stephen M. Gagola, Jr., Kent State University, Kent, OH. Let T_n denote the inverse image in R of the set of elements of finite order in $R/[R, nF]$, so $T_{n+1} \subseteq T_n$. The problem asks for a proof that $T_n/[R, nF]$ is an abelian subgroup of $R/[R, nF]$. We show that $T_n/[R, (n+1)F]$ is a *central* subgroup of $R/[R, (n+1)F]$ and hence is abelian. Since $T_n/[R, nF]$ is a homomorphic image of $T_n/[R, (n+1)F]$, the desired result follows.

Let $\{F_n\}$ denote the terms of the lower central series of F, defined recursively by $F_1 = F$ and $F_{n+1} = [F_n, F]$ for $n \ge 1$. The only property of free groups used here is that F/F_n is torsion-free for all n. Since $[R, nF] \subseteq [F, nF] = F_{n+1}$ and F/F_{n+1} is torsion-free, it follows that $T_n \subseteq R \cap F_{n+1}$, so $[R, T_n] \subseteq [R, F_{n+1}]$. To prove that $T_n/[R, (n+1)F]$ is central in $R/[R, (n+1)F]$ (and hence is an abelian group), it suffices to prove that $[R, F_{n+1}] \subseteq [R, (n+1)F]$.

We prove by induction on n that $[R, F_n] \subseteq [R, nF]$ for every n and every normal subgroup R. Equality holds when $n = 1$. For $n > 1$, we inductively assume that $[N, F_{n-1}] \subseteq [N, (n-1)F]$ for every normal subgroup N. By the Three Subgroups Lemma,

$$\big[\,[F_{n-1}, F], R\,\big] \subseteq \big[\,[R, F_{n-1}], F\,\big] \cdot \big[\,[F, R], F_{n-1}\,\big].$$

But

$$\big[\,[R, F_{n-1}], F\,\big] \subseteq \big[\,[R, (n-1)F], F\,\big] = [R, nF]$$

and

$$\big[\,[F, R], F_{n-1}\,\big] = \big[\,[R, F], F_{n-1}\,\big] \subseteq \big[\,[R, F], (n-1)F\,\big] = [R, nF].$$

Hence $[R, F_n] = \big[\,[F_{n-1}, F], R\,\big] \subseteq [R, nF]$.

Solved also by A. M. Gaglione & D. Spellman and the proposer.

Semi-unfriendly Sets

10343 [1993, 874]. *Proposed by David M. Bloom, Brooklyn College, CUNY, Brooklyn, NY.* Let us call a subset of $\mathbb{Z}$ *semi-unfriendly* (abbreviated *S-U*) if it contains no three consecutive integers. Let E_n denote the n element set $\{1, 2, \ldots, n\}$ and let

$$A(n, k) = \#\{\, S \subset E_n : \#S = k,\ S \text{ is } S\text{-}U\}$$
$$B(n, k) = \#\{\, S \subset E_n : \#S = k,\ S \text{ is } S\text{-}U \text{ and } E_n - S \text{ is } S\text{-}U\}.$$

Prove that $B(3n-1, n) = A(n+3, 3)$ for all $n \geq 1$.

Solution I by the late Raphael M. Robinson. $A(n+3, 3)$ is the number of 3-element subsets of E_{n+3} that are not composed of three consecutive integers. Since there are $n+1$ consecutive triples,

$$A(n+3, 3) = \binom{n+3}{3} - (n+1) = n(n+1)(n+5)/6.$$

To compute $B(3n-1, n)$, consider sequences $0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 3n$ such that $x_{i+1} - x_i \leq 3$ for $0 \leq i \leq n$. The condition requires that $E_{3n} - \{x_1, x_2, \ldots, x_{n+1}\}$ contains no consecutive triple. Since the sum of the $n+1$ differences is $3n$, the differences must all equal 3, except for three differences, each 2, or two differences, one 1 and one 2. Hence $x_1, \ldots, x_{n+1}$ also contains no consecutive triple. Choosing the locations for the differences other than 3 yields

$$B(3n-1, n) = \binom{n+1}{3} + (n+1)n = n(n+1)(n+5)/6.$$

Thus $A(n+3, n) = B(3n-1, n)$.

Composite solution II by Richard Holzsager, American University, Washington, DC and the proposer. Encode a set $T \subseteq E_n$ by a string $\sigma(T) = u_1 u_2 \ldots u_n u_{n+1}$, where $u_{n+1} = 0$ and otherwise $u_i = 1$ if $i \in T$ and $u_i = 0$ if $i \notin T$. Then T is *S-U* if and only if $\sigma(T)$ can be decomposed into substrings of the form 0, 10, and 110 (such a decomposition is unique). If $\sigma(T)$ uses a, b, c of the three types of substrings, respectively, then $n+1 = a + 2b + 3c$ and $\#(T) = b + 2c$. If we change each 0 to 110 and each 110 to 0 in the decomposition of $\sigma(T)$, we obtain $\sigma(T')$ for a *S-U* set T' of size $2a + b = 2(n+1) - 3\#(T)$. Furthermore, the length of $\sigma(T')$ is $3a + 2b + c = 3(n+1) - 4\#(T)$. This bijection proves that $A(n, k) = A(3n - 4k + 2, 2n - 3k + 2)$. In particular, $A(n+3, 3) = A(3n-1, 2n-1)$.

By complementation, $B(3n-1, n) = B(3n-1, 2n-1)$. As argued in Solution I, the complement of every semi-unfriendly $2n-1$-element subset of E_{3n-1} is also semi-unfriendly, so $B(3n-1, 2n-1) = A(3n-1, 2n-1)$. Hence $B(3n-1, n) = A(n+3, 3)$.

Editorial comment. Charles Lanski and Uday S. Gandbhir gave a more direct bijection from the set counted by $A(n+3,3)$ to the set counted by $B(3n-1,n)$: If a set is counted by $A(n+3,3)$, let its complement in E_{n+3} be $\{x_1,\dots,x_n\}$, where $x_1<\cdots<x_n$. Then $B(3n-1,n)$ counts the sets $\{4-x_1, 8-x_2,\dots,4n-x_n\}$, and the correspondence is bijective.

J. C. Binz generalized the result to *m-unfriendly* sets, which contain no consecutive m-tuple. Letting $A_m(n,k)$ and $B_m(n,k)$ count the m-unfriendly k-subsets of E_n and the m-unfriendly k-subsets of E_n whose complements are also m-unfriendly, he proved that $B(mn-1,n)=\binom{m+n}{m}-(n+1)=A(n+m,m)$.

Solved also by R. Barbara (Lebanon), D. Beckwith, J. Boutillon (France), S. Byrd, D. Callan, U. S. Gandbhir (Switzerland), D. S. Gunderson, R. D. Hurwitz, K. S. Kedlaya, N. Komanda, C. Lanski, G. Lord, O. P. Lossers (The Netherlands), A. Pedersen (Denmark), D. Wolfe, A. N. 't Woord (The Netherlands), Anchorage Math Solutions Group, and the NSA Problems Group (two solutions).

Cubic Polynomials from Curious Sums

10346 [1993, 951]. *Proposed by David Doster, Choate Rosemary Hall, Wallingford, CT.* Prove that, for all primes p,

$$\sum_{k=1}^{p-1}\left\lfloor\frac{k^3}{p}\right\rfloor=\frac{(p-2)(p-1)(p+1)}{4}; \qquad (A)$$

and

$$\sum_{k=1}^{M}\left\lfloor\sqrt[3]{kp}\right\rfloor=\frac{(3p-5)(p-2)(p-1)}{4}, \qquad (B)$$

where $M=(p-1)(p-2)$.

Solution to (A) by Manjul Bhargava (student), Harvard University, Cambridge, MA. For $1\le k\le p-1$, we have $k^3\not\equiv 0 \pmod p$ and $(p-k)^3\equiv -(k^3)\pmod p$, and therefore

$$\left(\frac{k^3}{p}-\left\lfloor\frac{k^3}{p}\right\rfloor\right)+\left(\frac{(p-k)^3}{p}-\left\lfloor\frac{(p-k)^3}{p}\right\rfloor\right)=1.$$

Hence,

$$\begin{aligned}\sum_{k=1}^{p-1}\left\lfloor\frac{k^3}{p}\right\rfloor&=\sum_{k=1}^{p-1}\frac{k^3}{p}-\sum_{k=1}^{p-1}\left(\frac{k^3}{p}-\left\lfloor\frac{k^3}{p}\right\rfloor\right)\\&=\frac{1}{p}\left(\frac{p(p-1)}{2}\right)^2-\frac{p-1}{2}=\frac{(p-2)(p-1)(p+1)}{4}.\end{aligned}$$

Solution to (B) by Ed Shapiro, Hanover, NH, and Lou Shapiro, Howard University, Washington, DC. Consider the set of lattice points $S=\{(n,m): 1\le n\le M, 1\le m\le p-1\}$. Since p does not divide m, we have $m^3\ne np$ for all $(n,m)\in S$, and hence the curve $y=\sqrt[3]{xp}$ contains no point of S. Now the sum $\sum_{k=1}^{M}\lfloor\sqrt[3]{kp}\rfloor$ counts the points of S below the curve while $\sum_{k=1}^{p-1}\lfloor k^3/p\rfloor$ counts the points of S to the left of the curve. Since $\#(S)=(p-2)(p-1)^2$, we use (A) to obtain

$$\sum_{k=1}^{M}\left\lfloor\sqrt[3]{kp}\right\rfloor=(p-2)(p-1)^2-\frac{(p-2)(p-1)(p+1)}{4}=\frac{(p-2)(p-1)(3p-5)}{4}.$$

Editorial comment. The selected solutions prove the result, observed by many solvers, that the formulas hold whenever p is square-free. More generally, for any integer p, let a be the number of positive integers less than p whose cube is divisible by p. Some readers showed

that (A) holds when p is any integer if the right side is augmented by $a/2$. Since this is half the number of lattice points on the curve $y = \sqrt[3]{xp}$ appearing in the selected solution to (B), the right side of (B) must also be increased by this amount. Other readers retained the assumption that p be prime and investigated the effect of replacing the exponent 3 by an arbitrary odd integer r. These generalizations were combined by K. S. Williams, who gave formulas involving Bernoulli numbers for the general sum of this type.

Solved also by A. Adelberg, J. Alvarez (Spain), J. Anglesio (France), R. Bagby, R. Barbara (Lebanon), A. Bergman, K. L. Bernstein, J. C. Binz (Switzerland), R. J. Chapman (U. K.), W. Chu (China), D. A. Darling, C. A. DeCarlucci, P. L. Douillet (France), J. S. Frame, S. M. Gagola Jr., U. S. Gandbhir (Switzerland), M. Getz, J. Greene, R. Holzsager, K. S. Kedlaya, M. J. Knight, H. K. Krishnapriyan, K.-W. Lau (Hong Kong), O. P. Lossers (The Netherlands), D. E. Manes, A. Nijenhuis, A. Pedersen (Denmark), R. M. Robinson, A. J. Rosenthal & D. E. Arias, K. Schilling, F. Schmidt, C. Schoen, M. Vowe (Switzerland), H. Widmer (Switzerland), K. S. Williams (Canada), A. N. 't Woord (The Netherlands), NSA Problems Group, and the proposer.

A 4-Angle Criterion for Concurrence

10348 [1993, 952]. *Proposed by Jiang Huanxin (student), FuDan University, ShangHai, China.* Let D, E, F be distinct points on the sides BC, CA, and AB respectively of $\triangle ABC$. Let $\alpha = \angle BDF$, $\beta = \angle FDA$, $\gamma = \angle ADE$, and $\delta = \angle EDC$. If AD, BE, and CF are concurrent and $\alpha/\beta = \delta/\gamma = m$ $(m \neq 1)$, prove that $\alpha = \delta$ and $\beta = \gamma$.

Solution by Nasha Komanda, Central Michigan University, Mt. Pleasant, MI. Use the proportions

$$\frac{\text{Area}(\triangle BDF)}{\text{Area}(\triangle ADF)} = \frac{|BF|}{|AF|} = \frac{|BD|\sin\alpha}{|AD|\sin\beta}$$

and

$$\frac{\text{Area}(\triangle CDE)}{\text{Area}(\triangle ADE)} = \frac{|CE|}{|AE|} = \frac{|CD|\sin\delta}{|AD|\sin\gamma}$$

to obtain

$$\frac{|BF|\cdot|AE|}{|AF|\cdot|CE|} = \frac{|BD|\sin\alpha\sin\gamma}{|CD|\sin\beta\sin\delta}.$$

By Ceva's Theorem, $|BF|\cdot|AE|\cdot|CD| = |AF|\cdot|CE|\cdot|BD|$. Therefore,

$$\frac{\sin\alpha}{\sin\beta} = \frac{\sin\delta}{\sin\gamma}. \tag{$*$}$$

If in addition $\alpha/\beta = \delta/\gamma = m$ $(m \neq 1)$, then $f(\beta) = f(\gamma)$ where $f(x) = \sin(mx)/\sin x$. The equality $\beta = \gamma$ (which implies $\alpha = \delta$) follows if we prove that f is strictly monotone on $(0, a)$ where $a = \pi$ if $m < 1$ and $a = \pi/m$ if $m > 1$.

Since $f'(x) = g(x)/\sin^2 x$ with $g(x) = m\cos(mx)\sin x - \cos x\sin(mx)$, it suffices to show that $g(x)$ does not change sign on $(0, a)$. However, $g'(x) = \left(1 - m^2\right)\sin(mx)\sin x$ and $g(0) = 0$. Thus, $g(x) > 0$ on $(0, a)$ if $m < 1$ and $g(x) < 0$ on $(0, a)$ if $m > 1$.

Editorial comment. The first part of the proof shows that formula $(*)$ is equivalent to AD, BE, and CF being concurrent when the "sides" of the triangle are taken to be line segments rather than extended lines. Other applications of a general concurrence criterion based on $(*)$ would be interesting.

The interval used in the second part of the proof is long enough to cover all intended geometric interpretations of the angles α, β, γ, and δ.

For $m \neq 1$, the conditions of the problem force the segment AD to be an altitude of $\triangle ABC$. Conversely, if AD is an altitude, $(*)$ reduces to $\tan\alpha = \tan\delta$, which is equivalent to $\alpha = \delta$. On the other hand, if AD is not an altitude, the general example of concurrent lines does not have $\alpha/\beta = \delta/\gamma$; but, for any D, examples with $m = 1$ are easily constructed. Since $(*)$ is then satisfied, this gives an example in which AD, BE and CF are concurrent.

Solved also by R. Barbara (Lebanon), R. J. Chapman (U. K.), A. Coffman, H. W. Guggenheimer, K. S. Kedlaya, Y.-H. Kiem (Korea), O. P. Lossers (The Netherlands), A. Nijenhuis, M. Vowe (Switzerland), and the proposer.

A Lemma of Dickson

10350 [1993, 952]. *Proposed by Borislav Lazarov, Sofia, Bulgaria.* Let M be a set of positive integers. Let P_M be the set of all primes that divide elements of M, and let L_M be the set of elements of M having no proper divisor in M. Show that P_M finite implies L_M finite.

Solution by the late Raphael M. Robinson. Let $P_M = \{p_1, p_2, \ldots, p_n\}$; use induction on n. If $n = 0$, then M is $\emptyset$ or $\{1\}$, and $L_M = M$. If $n = 1$, then $|L_M| = 1$. Suppose $n > 1$, and assume the result for $|P_M| < n$. Choose a fixed element $p_1^{r_1} p_2^{r_2} \cdots p_n^{r_n}$ of L_M. If $p_1^{s_1} p_2^{s_2} \cdots p_n^{s_n}$ is another element of L_M, then $s_k < r_k$ for some k. We need show only that the set of elements of L_M satisfying any one of these inequalities is finite. There are r_k choices for s_k, so it is sufficient to show that the set of elements of L_M with s_k fixed is finite. Let M' be the set of numbers x prime to p_k such that $p_k^{s_k} x \in M$. Then $|P_{M'}| < n$, so $L_{M'}$ is finite. Hence the set of elements of L_M with the prescribed s_k is also finite, since it is a subset of the set $p_k^{s_k} L_{M'}$ obtained by multiplying elements of $L_{M'}$ by $p_k^{s_k}$.

Editorial comment. Paul Erdős noted that the assertion of this problem appeared as Lemma B (an immediate corollary of Lemma A) in L. E. Dickson, Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors, *Amer. J. Math.* 35 (1913), 413–422 (or *Collected Mathematical Papers of Leonard Eugene Dickson*, Chelsea, 1975, Vol. 1, 349–358). Dickson gave two proofs of his Lemma A, one by induction along the lines of the selected solution and one by using the Hilbert Basis Theorem.

A generalization of Dickson's Lemma appeared as MONTHLY problem 4358 [1949, 480; 1952, 255]. While Problem 4358 contains the assertion of the present problem, its solution is somewhat more complicated.

Solved also by A. Adelberg, R. Barbara (Lebanon), D. Beckwith, K. L. Bernstein, M. Bhargava, P. Budney, D. Caccia, R. J. Chapman (U. K.), M. Dawes (Canada), P. Erdős (Hungary), K. Fabian & V. Sandor (Germany), S. M. Gagola Jr., R. Holzsager, K. S. Kedlaya, H. K. Krishnapriyan, O. P. Lossers (The Netherlands), R. MacDonald, A. Nijenhuis, V. Pambuccian, A. Pedersen (Denmark), A. Riese, K. Schilling, F. Schmidt, J. Simpson (Australia), M. Woltermann, A. N. 't Woord (The Netherlands), X. Xarles (Spain), NSA Problems Group, and the proposer.

The Ratio of Volume to Surface Area

10352 [1993, 952]. *Proposed by Yves Nievergelt, Eastern Washington University, Cheney, WA.* Let U be an open subset of $\mathbb{R}^n$ with smooth boundary ∂U contained in a ball of radius R. (**a**) For $n = 3$, show that $\mathrm{Vol}(U) \le R \cdot \mathrm{Area}(\partial U)/3$. (**b**) Generalize to arbitrary dimensions n.

Solution I by Richard Holzsager, The American University, Washington, DC. By the divergence theorem, if F is a vector field, then the integral over ∂U of the dot product of F with the outward normal ν is equal to the integral of $\nabla \cdot F$ over U. Placing the origin at the center of the given ball and taking F to be the radial vector field $F(\mathbf{p}) = \mathbf{p}$, we get $|F \cdot \nu| \le R$ and $\nabla \cdot F = 3$, giving the result.

The divergence theorem generalizes to any number of dimensions. Repeating this argument in n dimensions, we get $n\,\mathrm{Vol}(U) \le R\,\mathrm{Area}(\partial U)$ since $\nabla \cdot F = n$.

Solution II by Paul Sisson, Louisiana State University, Shreveport, LA. Let B_r^n be the ball of radius r and S_r^{n-1} the sphere of radius r in $\mathbb{R}^n$. The isoperimetric inequality is $\mathrm{Vol}(U)^{(n-1)/n} \le C_n\,\mathrm{Area}(\partial U)$ where $C_n = \mathrm{Vol}(B_r^n)^{(n-1)/n}/\mathrm{Area}(S_r^{n-1})$. Thus

$$\begin{aligned}\mathrm{Vol}(U) &\le C_n\,\mathrm{Vol}(U)^{1/n}\,\mathrm{Area}(\partial U)\\ &\le C_n\,\mathrm{Vol}(B_R^n)^{1/n}\,\mathrm{Area}(\partial U)\\ &= \left(\frac{\mathrm{Vol}(B_R^n)^{\frac{n-1}{n}}\,\mathrm{Vol}(B_R^n)^{\frac{1}{n}}}{\mathrm{Area}(S_R^{n-1})}\right)\mathrm{Area}(\partial U).\end{aligned}$$

Since $\mathrm{Vol}(B_R^n) = (R/n)\,\mathrm{Area}(S_R^{n-1})$, we have $\mathrm{Vol}(U) \le (R/n)\,\mathrm{Area}(\partial U)$ in $\mathbb{R}^n$.

Editorial comment. Both Richard Holzsager and Erik I. Verriest noted that the sphere in solution II need not contain U. It suffices to assume only that its volume is greater than or equal to $\mathrm{Vol}(U)$.

This problem is part of problem 6 of section 14 in H. Guggenheimer, *Applicable Geometry*, Krieger, 1977.

Solved also by R. J. Chapman (U. K.), H. W. Guggenheimer, M. S. Klamkin (Canada), T. C. Lim, O. P. Lossers (The Netherlands), G. Márton (Hungary), T. A. Murdoch, A. Nijenhuis, F. Schmidt, P. Szeptycki, E. I. Verriest (France), and the proposer.

Natural Linear Combinations

10354 [1994, 75]. *Proposed by Hassan Ali Shāh Ali, Tehran, Iran.* Determine the least natural number N such that, for all $n \ge N$, there exist natural numbers a, b with $n = \lfloor a\sqrt{2} + b\sqrt{3} \rfloor$.

Solution I by O. P. Lossers, University of Technology, Eindhoven, The Netherlands. There are two cases depending on whether 0 is considered to be a natural number. We first treat the case in which 0 is allowed. For every k, the numbers $\lfloor a\sqrt{2} + b\sqrt{3} \rfloor$ with $a+b = k$ (and $a, b \ge 0$) represent all integers in the range from $\lfloor k\sqrt{2} \rfloor$ to $\lfloor k\sqrt{3} \rfloor$, since $\sqrt{3} - \sqrt{2} \le 1$. For $k = 0, 1, 2$, this gives the intervals $[0, 0]$, $[1, 1]$, $[2, 3]$. Also for $k \ge 2$, we have $(k+1)\sqrt{2} - k\sqrt{3} < 1$. Hence every natural number is covered and $N = 0$ in this case.

If 0 is not considered a natural number, a similar analysis shows that $N = 3$.

Solution II by Albert Nijenhuis, University of Pennsylvania (Emeritus), Philadelphia, PA, and University of Washington, Seattle, WA. Assume that 0 is not a natural number. Then the minimal number representable in the form $\lfloor a\sqrt{2} + b\sqrt{3} \rfloor$ with natural numbers a, b is $\lfloor \sqrt{2} + \sqrt{3} = 3 \rfloor$. Thus $N \ge 3$.

Let $n \ge 3$ and set $m = \lfloor (n - \sqrt{3})/\sqrt{2} \rfloor$. Then

$$\sqrt{3} + m\sqrt{2} < n < \sqrt{3} + (m+1)\sqrt{2},$$

where equality cannot occur because of irrationality. If $\sqrt{3} + (m+1)\sqrt{2} < n + 1$, then $a = m + 1$ and $b = 1$ gives a representation of n. Otherwise, $\sqrt{3} + (m+1)\sqrt{2} > n + 1$, so that $n - (\sqrt{3} + m\sqrt{2}) < \sqrt{2} - 1$. Since $\sqrt{2} - 1 < 2(\sqrt{3} - \sqrt{2}) < 1$, it follows that

$$\sqrt{3} + m\sqrt{2} < n,\ \sqrt{3} + m\sqrt{2} + 2(\sqrt{3} - \sqrt{2}) = 3\sqrt{3} + (m-2)\sqrt{2} < n + 1,$$

so n is representable with $a = m-2$ and $b = 3$, provided that $m > 2$. Thus all $n > 3\sqrt{2}+\sqrt{3}$ are representable, giving $N \le 6$. In addition $4 = \lfloor 2\sqrt{2} + \sqrt{3} \rfloor$ and $5 = \lfloor 3\sqrt{2} + \sqrt{3} \rfloor$, so $N = 3$. Note that these representations require only $b = 1$ or $b = 3$.

Editorial comment. Most readers noted that this question has two answers depending on whether 0 is accepted as a natural number. Frank Schmidt noted that, for any natural number k, all sufficiently large integers have at least k representations in this form. Indeed, the number of representations of n is asymptotic to $n/\sqrt{6}$. This follows from results in G. H. Hardy & J. E. Littlewood, Some problems of Diophantine approximation, *Proc. London Math. Soc.*, (2) 20 (1922), 15–36 (or G. H. Hardy, *Collected Papers*, Vol. 1, Cambridge, 1966, 136–158). Patrick Dale McCray used a method similar to Solution II and found that the only properties used were that $\sqrt{2}$ is irrational and $0 < \sqrt{3} - \sqrt{2} < 1$.

Solved also by J. Alvarez (Spain), J. Anglesio (France), R. Barbara (Lebanon), K. L. Bernstein, R. E. Bernstein, S. Byrd, R. J. Chapman (U. K.), J. Christopher, D. A. Darling, Z. Franco, H. Gintis, R. Holzsager, K.-W. Lau (Hong Kong), J. H. Lindsey II, P. D. McCray, R. M. Robinson, F. Schmidt, M. Shemesh (Israel), P. Sisson, R. de la Vega (Colombia), M. Vowe (Switzerland), A. N. 't Woord (The Netherlands), The Ionasphere, NSA Problems Group, Westmont College Problem Solving Group, and the proposer.

REVIEWS

Edited by **Underwood Dudley**

Mathematics Department, DePauw University, Greencastle, IN 46135

Mathematics and Politics: Strategy, Voting, Power and Proof. By Alan D. Taylor, Springer-Verlag, 1995

Reviewed by **Samuel Merrill, III**

Mathematics has long been central in the pursuit of the physical sciences and of major significance in such fields as biology and economics. Yet its impact on political science—the early work of the Marquis de Condorcet in the 18th century, Duncan Black in the 1940's, and others notwithstanding—has been slow in coming. In the past three decades, however, a mathematical way of thinking has come to play a respected role in the study of politics, and in many ways represents a cutting edge of the field. At the same time the mathematics of politics has entered elementary mathematics textbooks, as mathematicians see politics as a source of fascinating problems and paradoxes.

Two of the important strands of this effort are (1) how to translate the preferences of individual members of an electorate over several alternatives into a coherent social preference—the theory of social choice—and (2) how to model conflict between two or more entities and determine appropriate strategies for the antagonists—game theory. Much of the *raison d'être* of social choice stems from the simple paradox of voting. Suppose three voters have transitive preference ranking *abc*, *bca*, and *cab*, respectively, for three alternatives labeled a, b, and c. Then a is preferred by a majority of two to one over b, b is similarly preferred over c, and in turn c over a, so that the principle of majority preference does not yield a transitive social ordering.

Duncan Black [1] provided a sufficient condition to avoid this paradox, called single-peakedness of preferences, i.e., the existence of an ordering of the alternatives such that a plot of any voter's preferences has a single peak. This idea, along with the work of Hotelling and Downs, led to the additional structure of a spatial model of electoral competition in which each voter (and candidate) is assumed to have an ideal point in a finite-dimensional issue space. The voter's utility for candidates is a declining function of distance from that ideal point (see, e.g., Enelow and Hinich [4]). If this model is one-dimensional (e.g., represents the familiar liberal/conservative scale), preferences are single-peaked and a transitive social order is defined. But with two or more dimensions (representing two or more issues), this coherence falls apart again. For example, in a two-dimensional model, place three candidates at the three roots of unity: $c_k = e^{2ki\pi/3}$, $k = 1, 2, 3$, and three voters at the same positions slightly rotated, say: $v_k = e^{i\pi/6}c_k$. Then c_1 is favored over c_2, c_2 over c_3, and c_3 over c_1, each by two voters to one.

The problems of transitivity of a social ordering parallel those of finding undominated candidate strategies, and in particular, equilibria. For a two-candidate contest in a one-dimensional spatial model, the location of the median voter

is an undominated strategy for a candidate to assume. In fact, it constitutes an equilibrium for both candidates, because any unilateral deviation from this stance risks losing votes. But in two or more dimensions, only a point that is the intersection of all median hyperplanes would constitute such an equilibrium; the existence of such a point is highly unlikely. Hence we can expect instability if candidates compete on more than one issue, or indeed as long as any candidate running behind can introduce and exploit a new issue. For three candidates there's no equilibrium at all even for one dimension; for four candidates in one dimension, there are two pairs of equilibrium positions if the voters are uniformly distributed; for more candidates, the situation gets progressively more complex (see Cox [**3**]).

Game theory—as a model of conflict—has played a significant role in economics and political science since the work of von Neumann and Morgenstern [**10**]. One of the chief rewards of applying the theory is to tease out rational behavior from seemingly irrational outcomes. For example, two-person, zero-sum games always have a solution, but in general that solution requires each player to choose probabilistically among two or more strategies. Thus, in repeated play, a player—e.g., a political candidate—may appear indecisive while following a rational strategy. Many conflicts—although often modeled as zero-sum games—may better be represented as variable-sum games, in which interests are not diametrically opposed, but rather all players can do better (or worse) simultaneously. A basic solution concept is a Nash equilibrium—a pair of strategies by the two players from which neither can benefit by a unilateral deviation. The choice of the median voter's position described above is such a solution.

Real conflicts, however,—whether they be between politicians or polities—usually involve a sequence of moves and counter-moves. This characteristic has led to the theory of *moves* (Brams [**2**]), under which players are assumed to alternate with one another in choosing strategies (move or stay put) until neither changes strategies. A player is assumed to move only if s(he) expects the final outcome would be better if both players play rationally. These assumptions often lead to strategies that appear more reasonable than those predicted by classic game theory.

Alan Taylor's objective in *Mathematics and Politics: Strategy, Voting, Power and Proof* is to make available some serious mathematical ideas and techniques arising from problems in political science and conflict resolution to a student audience with little or no mathematical background. His is a teaching book: he is ever aware of what his readers do not know but can be taught when needed in order to develop the main ideas. The book is full of side comments to shepherd those students who may have misleading thoughts back into the fold.

This effective style is illustrated well in the chapters on voting power. Attempts to give disparate influence to voting units through weighted voting or other means are shown to apportion power in unexpected ways. In fact, power itself can be defined in several different ways. Cognizant of the beginning student, Taylor describes how a yes-no voting system can be specified by winning and losing subsets of voters called coalitions and in great detail explains the construction of the Shapley-Shubik and Banzhaf indices of voting power. But he also introduces—without recourse to unwieldy notation—a series of paradoxes: the new-member and donation paradoxes—under both of which the Banzhaf index suffers (Felsenthal and Machover [**5**])—and the bicameral paradox, which infects the Shapley-Shubik index (Felsenthal, Machover, and Zwicker [**6**]). The recent discovery of these paradoxes lets the student know that research in this area—although

accessible to the novice—is alive and active and that, furthermore, the jury is still out concerning the evaluation of indices. In fact, many of the more significant theorems portrayed in the book have been discovered only in the last ten years, often by the author himself in collaboration with his colleague, William Zwicker at Union College.

For example, a yes-no voting system is weighted if and only if it is *trade robust*, i.e., if an arbitrary exchange of players among several winning coalitions leaves at least one of the coalitions winning. A less restrictive condition is *swap-robustness*, which applies only to one-for-one exchanges. The U.S. federal system (Congress and the President) is not even swap robust (a House member and a Senator cannot in general be swapped); the amendment procedure for the Canadian Constitution —though swap robust—is not trade robust. Hence neither can be specified as a simple weighted system (with weights for the players and a quota). But any yes-no voting system can be specified via a finite collection of weighted voting systems each with the same set of voters such that the common winning coalitions constitute the winning coalitions of the original system. The minimum number of such systems needed defines the dimension of the yes-no voting system. The U.S. federal system and the Canadian Constitution each have dimension two. Alternatively, we may express each yes-no voting system as a Cartesian product of n weighted voting systems, where n is, of course, the dimension.

The same concern for the beginning student pervades the chapter on social choice (cf. Straffin's *Theory of Voting* [**8**], and Merrill's *Making Multicandidate Elections More Democratic* [**7**]). Attempts to extend the majoritarian principle to multicandidate elections lead to one paradox after another and eventually to Arrow's impossibility theorem. Taylor gives a series of formal proofs of properties such as the Condorcet winner criterion and monotonicity for various voting systems, each followed by a brief synopsis of the proof, which makes memorable the crux of the argument. Although a minor point in a generally excellent book, this makes the formal proofs seem unnecessary—almost pedantic. Should they be included? The chapter ends with the proof of a new result: no social choice system for three or more alternatives can satisfy both independence of irrelevant alternatives and the Condorcet winner criterion.

This thread is pursued later in the book—after the student has gained greater sophistication. Taylor proves May's theorem that the only procedure for two alternatives that is anonymous (invariant under permutation of voters), neutral (invariant under permutation of candidates), and monotone that produces a single winner for an odd number, n, of voters is majority rule. If ties are allowed, a quota system is obtained where $n/2 < q \leq n + 1$. He then proceeds to the classic impossibility theorem of Arrow that the only social welfare function for three or more candidates satisfying the Pareto condition, independence of irrelevant alternatives, and monotonicity is a dictatorship. To wrap up, he proves theorems of Black and of Sen, each of which provide sufficient conditions on the coherence of preferences to guarantee a transitive social ordering.

On the topic of game theory, Taylor shows the student how political conflict—whether personal or international—can often be modeled by variable-sum games such as Prisoners' Dilemma or Chicken. Prisoners' Dilemma abstracts the frustrations at all levels of society between individual striving and cooperative agreement. The game of Chicken can embody the idea of deterrence; Taylor uses it to model two interpretations of the Cuban missile crisis. The "dollar auction," with which Taylor introduces his book, provides a rationale for the seemingly irrational behavior inherent in escalation. His treatment of applied game theory

could serve as an introduction to Straffin's undergraduate text, *Game Theory and Strategy* **[9]**.

Alan Taylor's book is carefully crafted. He is ever aware of his audience, but relentlessly presses the beginning student to understand more and more ideas. The text is appropriate for bright, intellectually motivated, but mathematically untrained, undergraduates, who are provided with the opportunity to experience a significant frontier of mathematics.

REFERENCES

1. Black, Duncan. (1958). *The Theory of Committees and Elections*. Cambridge: Cambridge University Press.
2. Brams, Steven J. (1994). *Theory of Moves*. Cambridge: Cambridge University Press.
3. Cox, Gary. (1987). Electoral Equilibrium under Alternative Voting Institutions. *American Journal of Political Science*, 31:82–108.
4. Enelow, James, and Melvin Hinich. (1984). *The Spatial Theory of Voting*. Cambridge: Cambridge University Press.
5. Felsenthal, D. S., and Moshé Machover. (1994). Postulates and paradoxes of relative voting power —a critical reappraisal. *Theory and Decision* 38:195–229.
6. Felsenthal, D. S., Moshé Machover, and William Zwicker. (1995). The Bicameral Postulates and Indices of *a priori* Relative Voting Power. Presented at the annual meeting of the Public Choice Society, Long Beach, CA, March 24–26.
7. Merrill, Samuel III. (1988). *Making Multicandidate Elections More Democratic*. Princeton: Princeton University Press.
8. Straffin, Philip. (1980). *Topics in the Theory of Voting*. Boston: Birkhauser.
9. Straffin, Philip. (1993). *Game Theory and Strategy*. Washington, D.C.: The Mathematical Association of America.
10. von Neumann, John, and Oskar Morgenstern. (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Department of Mathematics and Computer Science
Wilkes University
Wilkes-Barre, PA 18766
smerrill@wilkes.edu

Ever wonder why certain shampoos have just the right consistency, so they flow from the bottle but not through your fingers? Or why a particular soft drink seems to hit the spot, while another may not? Or how cyclists could shatter 21 speed records at Atlanta's Summer Olympic Games?

The answer in each case: mathematics.

Most ordinary Americans may hate doing math, but they are increasingly subject to its influence, whether they know it or not. Experts in a growing number of fields are relying on calculation rather than estimation. Some textile designers use geometry to create pleasing new designs. Many companies have discarded rule of thumb in favor of sophisticated statistical tools to decide what products to place on supermarket shelves. Even musicians use math equations to generate inventive new sounds.

Technology, *The Wall Street Journal*, September 25, 1996, p. B-1.

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of
Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

Recreational Mathematics, S. *3-D Geometric Origami: Modular Polyhedra.* Rona Gurkewitz, Bennett Arnstein. Dover, 1995, iv + 73 pp, $6.95 (P). [ISBN 0-486-28863-3] Step-by-step instructions and clear diagrams for constructing over 50 modular polyhedra-based models. JNC

Education, P. *Mathematical Education of Engineers.* Eds: L.R. Mustoe, S. Hibberd. Inst. of Math. & Its Applic. Conf. New Ser., No. 57. Clarendon Pr, 1995, xxiii + 383 pp, $120. [ISBN 0-19-851191-4] Proceedings of a 1994 conference at Loughborough University.

History, P. *A New Branch of Mathematics.* Hermann Grassmann. Transl: Lloyd C. Kannenberg. Open Court, 1995, xvi + 555 pp, $32.95 (P). [ISBN 0-8126-9275-6] Translation of Grassmann's "Die lineale Ausdehnungslehre," "Geometric Analysis," and other selected papers. LC

History, P, L. *Vita Mathematica: Historical Research and Integration with Teaching.* Ed: Ronald Calinger. MAA, 1996, xii + 359 pp, $34.95 (P). [ISBN 0-88385-097-4] Papers on the history of mathematics and its integration with the teaching of mathematics. Topics range from historical surveys (for example, "The Combinatorics and Induction in Medieval Hebrew and Islamic Mathematics," by Katz) to pedagogy ("History of Mathematics and the Teacher," by Heiede). Valuable resource. LC

History, P. *Of Men and Numbers: The Story of the Great Mathematicians.* Jane Muir. Dover, 1996, 249 pp, $7.95 (P). [ISBN 0-486-28973-7] Republication of the 1961 Dodd, Mead & Co. edition.

History, P. *Courant.* Constance Reid. Copernicus (Imprint: Springer-Verlag), 1996, 318 pp, $15 (P). [ISBN 0-387-94670-5] Republication of *Courant in Göttingen and New York* (TR, January 1987).

History, P. *Hilbert.* Constance Reid. Copernicus (Imprint: Springer-Verlag), 1996, 228 pp, $15 (P). [ISBN 0-387-94674-8] Paperback republication of the original 1970 edition (TR, June–July 1970; Extended Review, May 1971).

Number Theory, P. *Fundamentals of Number Theory.* William J. LeVeque. Dover, 1996, vii + 280 pp, $8.95 (P). [ISBN 0-486-68906-9] Republication of the 1977 Addison-Wesley edition.

Number Theory, T*(13–14: 1). *Introduction to Number Theory.* Peter D. Schumer. PWS, 1996, xi + 287 pp. [ISBN 0-534-94626-7] A very well-written introduction to number theory with good examples and problems, and many historical asides. Includes material on factorization and primality testing, continued fractions, partition theory, and an introduction to analytic number theory. DB

Group Theory, P. *Groups, Difference Sets, and the Monster.* Eds: K.T. Arasu, *et al.* Ohio St. Univ. Math. Res. Inst., V. 4. Walter de Gruyter, 1995, xiii + 461 pp, DM 198. [ISBN 3-11-014791-2] Proceedings of a 1993 special research quarter at The Ohio State University.

Algebra, P. *Tight Closure and Its Applications.* Craig Huneke. CBMS Reg. Conf. Ser. in Math.,

No. 88. AMS, 1996, ix + 137 pp, $29 (P). [ISBN 0-8218-0412-X]

Algebra, P. *Cogroups and Co-rings in Categories of Associative Rings.* George M. Bergman, Adam O. Hausknecht. Math. Surv. & Mono., V. 45. AMS, 1996, ix + 388 pp, $79. [ISBN 0-8218-0495-2]

Calculus, S(13). *Discovering Calculus with the Graphing Calculator.* Mary Margaret Shoaf-Grubbs. Wiley, 1996, xiv + 204 pp, $23.95 (P). [ISBN 0-471-00974-1] Designed to accompany a standard calculus text. A few introductory pages are followed by a series of labs each of which includes descriptions of applicable calculator features. Illustrated by screen examples from the TI–82 calculator. JNC

Calculus, S*(13). *CalcLabs with Mathematica.* Nancy R. Blachman, *et al.* Brooks/Cole, 1996, xvi + 245 pp, $20.25 (P). [ISBN 0-534-34086-5]; *CalcLabs with Maple V.* Albert Boggess, *et al.* Brooks/Cole, 1995, xiii + 229 pp, $20.95 (P). [ISBN 0-534-25590-6] Designed for a first-year calculus class with 1 hour of lab per week. The first 13 or 14 chapters contain user-friendly introductions to commonly-used commands as tools for solving calculus problems, and conclude with exercises. The last two chapters contain labs and longer student projects. The *Mathematica* version assumes Version 2.2 and use of the Notebook Front End; the *Maple V* Version assumes Release 3. JNC

Complex Analysis, T(18), S. *Entire and Meromorphic Functions.* Lee A. Rubel. Universitext. Springer-Verlag, 1996, viii + 187 pp, $39 (P). [ISBN 0-387-94510-5] Begins with a clear and concise treatment of Nevalinna theory. Develops the Rubel-Taylor method of Fourier series analysis. Presents Pólya's theory of the Borel Transform and Buck's theory of integer valued entire functions. The writing is sparse: lemma-theorem-corollary format. Contains a small bibliography and no exercises. TAV

Partial Differential Equations, P. *Partial Differential Equations of Mathematical Physics and Integral Equations.* Ronald B. Guenther, John W. Lee. Dover, 1995, xii + 562 pp, $17.95 (P). [ISBN 0-486-68889-5] Republication, with corrections, of the 1988 Prentice Hall edition. Includes a new section on "Solutions and Hints to Selected Problems."

Partial Differential Equations, P. *A Practical Guide to Pseudospectral Methods.* Bengt Fornberg. Mono on Appl. & Computat. Math., V. 1. Cambridge Univ Pr, 1996, x + 231 pp, $54.95. [ISBN 0-521-49582-2] Pseudospectral methods are important in several applications areas (e.g., computational fluid dynamics, wave motion). Explains how, when, and why these methods work. AO

Dynamical Systems, T(15–17), S, L. *Oscillations in Planar Dynamic Systems.* Ronald E. Mickens. Ser. on Adv. in Math. for Appl. Sci., V. 37. World Scientific, 1996, xii + 319 pp, $48. [ISBN 981-02-2292-0] Complete revision of *Introduction to Nonlinear Oscillations* (TR, April 1982). New chapters discuss method of harmonic balance, and a general procedure for two coupled first-order differential equations based on Hopf bifurcation theorem and averaging. Extensive bibliography. DH

Numerical Analysis, T(16–17: 2), L. *A First Course in the Numerical Analysis of Differential Equations.* Arieh Iserles. Texts in Appl. Math. Cambridge Univ Pr, 1996, xvii + 378 pp, $27.95 (P); $74.95. [ISBN 0-521-55655-4; 0-521-55376-8] Written for mathematics (rather than engineering) students. Covers the solution of ODEs by multistep and Runge-Kutta methods; finite difference and finite element methods for the Poisson equation; basic methods for parabolic and hyperbolic PDEs. AO

Numerical Analysis, P, L*. *Numerical Methods for Least Squares Problems.* Åke Björck. SIAM, 1996, xvii + 48 pp, $47.50 (P). [ISBN 0-89871-360-9] A comprehensive and up-to-date treatment that includes many recent developments. In addition to basic methods, it covers methods for modified and generalized least squares problems, and direct and iterative methods for sparse problems. AO

Numerical Analysis, P. *Lectures on Finite Precision Computations.* Françoise Chaitin-Chatelin, Valérie Frayssé. SIAM, 1996, xvi + 235 pp, $44.50 (P). [ISBN 0-89871-358-7] Addresses how finite precision affects the convergence in practice of numerical methods that are known to converge theoretically. DH

Functional Analysis, T(18), P. *Linear Functional Equations: Operator Approach.* Anatolij Antonevich. Transl: Victor Muzafarov, Andrei Iacob. Oper. Theory: Adv. & Applic., V. 83. Birkhäuser Boston, 1996, viii + 179 pp, $123. [ISBN 0-8176-2931-9] A unified approach to the investigation of a general class of functional equations based on the examination of functional operators and Banach algebras generated by them. Uses methods involving dynamical systems, operator algebras, and pseudodifferential operators. SA

Functional Analysis, P. *Elementary Functional Analysis.* Georgi E. Shilov. Transl: Richard A. Silverman. Dover, 1996, vii + 334 pp, $10.95 (P). [ISBN 0-486-68923-9] Republication, with corrections, of the 1974 MIT Press edition (Volume 2 of *Mathematical Analysis*).

Analysis, P. *Elementary Real and Complex Analysis, Revised English Edition.* Georgi E. Shilov. Transl. & Ed: Richard A. Silverman. Dover, 1996, xi + 516 pp, $12.95 (P). [ISBN 0-486-68922-0] Republication, with corrections, of the 1974 MIT Press edition (Volume 1 of *Mathematical Analysis*).

Analysis, P, L. *Padé Approximants, Second Edition.* George A. Baker, Jr., Peter Graves-Morris. Ency. of Math. & Its Applic., V. 59. Cambridge Univ Pr, 1996, xiv + 746 pp, $110. [ISBN 0-521-45007-1] Incorporates many new results and a new chapter on multiseries approximants. (*First Edition*, TR, November 1982.) AO

Analysis, P*, L.** *The World According to Wavelets: The Story of a Mathematical Technique in the Making.* Barbara Burke Hubbard. AK Peters, 1996, xix + 264 pp, $34. [ISBN 1-56881-047-4] An accessible and well-written book about wavelets for non-mathematicians. The first half recounts the development of this field of mathematics and contains (almost) no formulas. The second half ("Beyond Plain English") is a collection of articles that provide an elementary introduction to wavelets. AO

Analysis, P. *Lecture Notes in Control and Information Sciences–213: General Hybrid Orthogonal Functions and their Applications in Systems and Control.* Amit Patra, Ganti Prasada Rao. Springer-Verlag, 1996, xx + 118 pp, $43 (P). [ISBN 0-540-76039-3]

Analysis, P. *Potential Theory and Degenerate Partial Differential Operators.* Ed: Marco Biroli. Kluwer Academic, 1995, 184 pp, $99. [ISBN 0-7923-3596-1] Proceedings of a 1994 conference in Parma, Italy. Partially reprinted from *Potential Analysis*, V. 4 (1995).

Analysis, T(15–17: 1), P. *Linear Difference Equations with Discrete Transform Methods.* Abdul J. Jerri. Math. & Its Applic., V. 363. Kluwer Academic, 1996, xxi + 439 pp, $199. [ISBN 0-7923-3940-1] Tools for studying and solving ordinary linear difference equations. Covers, in addition to traditional techniques, the use of discrete Fourier transforms for solving boundary value problems. AO

Algebraic Geometry, P. *Abelian Functions: Abel's Theorem and the Allied Theory of Theta Functions.* H.F. Baker. Cambridge Univ Pr, 1995, xxxv + 684 pp, $39.95 (P). [ISBN 0-521-49877-5]

Differential Geometry, T(18: 1), P. *Basic Concepts of Synthetic Differential Geometry.* René Lavendhomme. Texts in Math. Sci., V. 13. Kluwer Academic, 1996, xv + 320 pp, $159. [ISBN 0-7923-3941-X] Introduction to synthetic differential geometry, an approach to differential geometry that uses infinitesimal elements (objects whose squares are 0) and intuitionist logic. JO

Geometry, S*, P, L**.** *Beyond the Third Dimension: Geometry, Computer Graphics, and Higher Dimensions.* Thomas F. Banchoff. Scientific American Library, 1996, ix + 211 pp, $19.95 (P). [ISBN 0-7167-6015-0] Paperback edition of the highly acclaimed volume by the guru of higher dimensions. New computer graphics illustrations enhance what was already a "visually rich and intellectually enriching" portrait of dimensions. (1990 hardcover edition, TR, August–September 1990; Extended Review, August–September 1991.) JNC

Topology, S*, P. *Counterexamples in Topology.* Lynn Arthur Steen, J. Arthur Seebach, Jr. Dover, 1995, xi + 244 pp, $8.95 (P). [ISBN 0-486-68735-X] Republication of the 1978 *Second Edition* originally published by Springer-Verlag (TR, January 1979).

Topology, P. *Lectures on Spaces of Nonpositive Curvature.* Werner Ballmann. DMV Sem., Band 25. Birkhäuser Boston, 1995, v + 112 pp, $32 (P). [ISBN 0-8176-5242-6]

Optimization, T(18: 1), P. *Modified Lagrangians and Monotone Maps in Optimization.* E.G. Golshtein, N.V. Tretyakov. Transl: N.V. Tretyakov. Ser. in Disc. Math. & Optim. Wiley, 1996, ix + 438 pp, $72.95. [ISBN 0-471-54821-9] Theory and applications of modified Lagrangian functions. Focuses on traditional convex programming and monotone maps. Applications include numerical algorithms for the general convex programming problem, decomposition, economic modeling, and nonconvex local constrained optimization. AO

Game Theory, T(13–16: 1), P*, L. *Fair Division: From Cake-Cutting to Dispute Resolution.* Steven J. Brams, Alan D. Taylor. Cambridge Univ Pr, 1996, xiv + 272 pp, $54.95; $18.95 (P). [ISBN 0-521-55390-3; 0-521-55644-9] Presents criteria for fairness, the latest constructive procedures for a fair division (e.g., envy-free allocation) of goods, and real-life applications. DH

Stochastic Processes, P. *Discrete-Time Markov Control Processes: Basic Optimality Criteria.* Onésimo Hernández-Lerma, Jean Bernard Lasserre. Applic. of Math., V. 30. Springer-Verlag, 1996, xiv + 216 pp, $54.95. [ISBN 0-387-94579-2] Markov control processes are standard tools in a wide variety of settings, from fish hatcheries to portfolio management. This book looks at the theoretical underpinnings and gives a solid treatment of the theory. Extensive bibiolography. TAV

Elementary Statistics, T(13: 2). *The New Statistical Analysis of Data.* T.W. Anderson, Jeremy D. Finn. Springer-Verlag, 1996, xxi + 712 pp, $59.95. [ISBN 0-387-94619-5] Updated and somewhat more elementary version of Anderson and Sclove's *The Statistical Analysis of Data* (*First Edition*, TR, December 1978). Changes from the 1986 *Second Edition* include additional chapters on descriptive measures and probability distributions, omission of the chapter on multiple regression, and a change to SPSS as the computer package of choice. RSK

Mathematical Computing, C, L. *The Mathematica Book, Third Edition.* Stephen Wolfram. Cambridge Univ Pr, 1996, 1395 pp, $59.95; $44.95 (P). [ISBN 0-521-58889-8; 0-521-58888-X] User guide and reference manual for *Mathematica 3.0*, the most recent version of this software. (1988 Addison-Wesley edition, TR, October 1988; Extended Review, November 1989.) AO

Mathematical Computing, P*. *The Maple Handbook: Maple V Release 4.* Darren Redfern. Springer-Verlag, 1996, 495 pp, $29 (P). [ISBN 0-387-94538-5] Reference tool. Brief entries for each command are organized in subject area categories (calculus, linear algebra, combinatorics, number theory, etc.). AO

Computer Science, P, L. *Practical UNIX and Internet Security, Second Edition.* Simson Garfinkel, Gene Spafford. O'Reilly & Associates, 1996, xxix + 971 pp, $39.95 (P). [ISBN 1-56592-148-8]

Applications (Fluid Mechanics), P. *Navier-Stokes Equations and Related Nonlinear Problems.* Ed: A. Sequeira. Plenum Pr, 1995, ix + 406 pp, $115. [ISBN 0-306-45118-2] Proceedings of a 1994 conference in Funchal, Portugal.

Applications (Fluid Mechanics), P. *Annual Review of Fluid Mechanics, V. 28, 1996.* Eds: John L. Lumley, Milton van Dyke, Helen L. Reed. Annual Reviews, 1996, x + 598 pp, $52. [ISBN 0-8243-0728-3]

Applications (Fluid Mechanics), P. *Computational Methods for Fluid Dynamics.* J.H. Ferziger, M. Perić. Springer-Verlag, 1996, xiv + 356 pp, $49.50 (P). [ISBN 3-540-59434-5] An overview of commonly used methods including direct and large eddy simulation of turbulence, multigrid methods, parallel computing, moving grids, and free surface flows. AO

Applications (Physics), T(18), P. *Evolution Processes and the Feynman-Kac Formula.* Brian Jefferies. Math. & Its Applic., V. 353. Kluwer Academic, 1996, ix + 235 pp, $125. [ISBN 0-7923-3843-X] The evolution of a physical system can often be described in terms of a semi-group of linear operators. Observations may be modelled by a spectral measure. A combination of these basic objects produces a family of operator-valued set functions, by which perturbations of the evolution are represented as path integrals. Integration theory in vector spaces is a central topic of this work. SA

Applications (Physics), P. *Angular Momentum in Quantum Mechanics.* A.R. Edmonds. Landmarks in Physics. Princeton Univ Pr, 1996, viii + 146 pp, $39.50 (P). [ISBN 0-691-07912-9] Republication of the 1974 corrected printing of the *Second Edition.*

Applications (Physics), P. *General Theory of Relativity.* P.A.M. Dirac. Landmarks in Physics. Princeton Univ Pr, 1996, viii + 71 pp, $10.95 (P). [ISBN 0-691-01146-X] Republication of the 1975 Wiley edition.

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–212: Formal Specification and Synthesis of Procedural Controllers for Process Systems.* Arturo Sanchez. Springer-Verlag, 1996, xxiv + 221 pp, $54 (P). [ISBN 3-540-76021-0]

Applications. *Monitoring a Comprehensive Test Ban Treaty.* Eds: Eystein S. Husebye, Anton M. Dainty. NATO ASI Ser. E, V. 303. Kluwer Academic, 1996, xxv + 836 pp, $349. [ISBN 0-7923-3811-1]

Applications, P, L*. *The Algorithmic Beauty of Plants.* Przemyslaw Prusinkiewicz, Aristid Lindenmayer. Springer-Verlag, 1996, xii + 228 pp, $29.95 (P). [ISBN 0-387-94676-4] Computer graphics techniques for modeling plant development and plant shapes. Emphasizes use of Lindenmayer systems. Many color plates. AO

Reviewers

SA: Steve Abbott, St. Olaf; DB: David Bressoud, Macalester; JNC: Judith N. Cederberg, St. Olaf; LC: Laura Chihara, St. Olaf; DH: Deanna Haunsperger, Carleton; RSK: Richard S. Kleber, St. Olaf; JO: Jeff Ondich, Carleton; AO: Arnold Ostebee, St. Olaf; TAV: Theodore A. Vessey, St. Olaf.

THE AUTHORS

BRUCE POURCIAU holds the usual degrees (B.A. from Brown, Ph.D. from UC San Diego under Hubert Halkin) and since 1976 has held the usual positions (assistant, associate, professor of mathematics) at Lawrence University, overlooking the serene Fox River in Appleton, Wisconsin. His mathematical interests include optimization theory, various blends of topology and analysis, the philosophy of mathematics, especially intuitionism, and the history of mathematics, especially Newton's *Principia*. When he isn't playing tennis, photographing nature, listening to Mozart, or reading mysteries, you will find him involved with his more important interests—his wife, Nancy, and their three children, two of whom now attend the usual colleges.

JACK ROGERS received his B.A., M.S., and Ph.D. degrees from the University of Texas at Austin, the last in 1966. He then taught at Emory University, moving to Auburn University in 1973, where he is Professor of Mathematics. His research interests are in numerical analysis, primarily in optimization.

ANDREW VINCE has lived in San Francisco and Stanford, CA, Cambridge and Woods Hole, MA, at the mouth of the Susitna River, AK, New Brunswick, ME, Ann Arbor, MI (where he received his Ph.D. in 1981), Zomba, Malawi, and Kampala, Uganda (thanks to Fulbright grants), and now resides in Gainesville, FL, where he is on the mathematics faculty at the University of Florida. His research interests are in combinatorics, graph theory, discrete geometry, polytopes, and tilings. According to his ten year old son, he is OK at math and cycling, but not so good at shooting hoops.

JOHN BAKER occupies the bucket seats of classical analysis and functional equations at the University of Waterloo where he is the resident expert on bawdy songs and sea chanties. In his mid-fifties and a life-long athletic supporter, Baker also likes eating, woodworking, perambulating, and people who encourage him to sing and play guitar. With Luddite leanings and curmudgeon characteristics, his major irritants include educationists, political correctness, mission statements, and voice mail. His greatest blessings are his wife Donna, son Sean, and daughter Laurel.

REUBEN HERSH studied partial differential equations at the Courant Institute in New York. He taught at several post-secondary institutions in the U.S. and Mexico. With Philip J. Davis, he wrote *The Mathematical Experience* and *Descartes' Dream*. On his own, he wrote the forthcoming *What Is Mathematics, Anyways?*. He is retired and lives in Santa Fe, New Mexico.

SAMUEL MERRILL, III, received a Ph.D. in mathematics from Yale University and an M.S. in statistics from the Pennsylvania State University. His current research involves mathematical and statistical modeling, particularly in voting procedures and behavior. This interdisciplinary interest was originally sparked by the 1973 MAA Summer Seminar on Mathematics and the Behavioral Sciences. He is the author of *Making Multicandidate Elections More Democratic* (1988: Princeton University Press) and has published in journals ranging from the *American Political Science Review* to the *Journal of the American Statistical Association.*

EDITOR'S ENDNOTES

Every five years or so, one editorial cycle ends and another begins. More than a year ago, members of the Editorial Board and staff listed inside the front cover began to solicit, read, select, and edit the articles and features for this issue and its 49 successors. We were ably tutored and generously nurtured by John Ewing and his Editorial Board, who helped us master the traditions and mechanics of the vigorous and healthy enterprise they have passed on to us.

We are fully committed to the MONTHLY's historic mission of publishing high-quality exposition of mathematics in order to advance and serve the broad spectrum of collegiate mathematics. As authors, we know that good expository writing about deep ideas for a general mathematical audience can be hard work, but as readers and teachers we know that it is well worth the effort. We know that our efforts have been successful when an article is interesting enough that our nonspecialist colleagues are willing to read it in bed, talk about it at coffee hour, lift a tidbit from it to present in class, or recommend it to a bright undergraduate for independent reading.

To ensure an interesting variety of articles in each issue, the Editorial Board is actively soliciting articles in history and biography, statistics, computer science, modern applied mathematics, and mathematics education. In the latter area, we are especially interested in submissions that: encourage communications between mathematicians and mathematics educators, and between the mathematics community and client disciplines; encourage mathematicians to reflect on their own teaching; and share applicable results from mathematics education research. Naturally, we expect the steady flow of articles from core mathematical areas to continue while we give extra encouragement to areas from which we now see only a trickle.

What qualities distinguish the few article submissions that are published (53 last year) from the many hundreds that are not? Novelty of ideas is neither necessary for acceptance nor sufficient for rejection. *Interesting* articles that present well-known ideas are welcome at the MONTHLY. *Interesting* articles that successfully present the fruits of current research to our broad audience of mathematical readers are welcome, too; papers addressed to experts at the frontier should be submitted to appropriate specialized journals. Award-winning MONTHLY articles have high expository quality, which means much more than grammar, punctuation, and spelling. Somehow they find an attractive way to invite the reader to begin, keep interest high with well-chosen examples and figures, illustrate key issues via artfully-chosen special cases, and reward the reader's active engagement by informing, enriching, and even entertaining. Clarity of exposition and broad appeal are always more important in a *Monthly* article than originality of the material or generality of the results.

There is no ideal length for a MONTHLY article, which could be quite short if that is appropriate to the material. However, an upper bound of at most 60 pages for articles in each issue means that every decision to publish a very long article is necessarily a decision to limit the variety of articles in an issue, and we are always reluctant to do that.

MONTHLY Notes are not short articles. In addition to insisting that Notes be readily accessible to a broad mathematical audience, we expect them to be mathematical gems that demonstrate significant originality in results, proof, or viewpoint.

And now a few words about technology. We are experimenting with new ways to communicate with readers and authors, and invite you to visit the MONTHLY's section of MAA Online at http://www.maa.org/. There you will find tables of contents of forthcoming issues, brief descriptive summaries of articles, and information for authors about submitting papers, providing TeX source files, and achieving clarity in mathematical writing.

Like all the Editors of the 36 years of MONTHLY volumes in my bookcase, I encourage you to write me about anything concerning our journal that you think is important. My predecessors tell me to expect very little mail from readers, but perhaps the convenience of

email will stimulate a flow of comments and reactions that can beneficially inform the choices we make daily in the editorial office.

Abe Shenitzer has suggested the following clarification of the lead paragraph on noncommutative ring theory on p. 418 of the May, 1996 issue (Vol. 103):

> Hamilton's "physical" motivation was to define an algebra of triples that would do for rotations in 3-space what the complex numbers do for rotations in the plane. Having failed in this task, he turned to quadruples of reals and created the algebra of quaternions. The quaternions did, in fact, yield the required computing tool for rotations in 3-space.

Gary White wrote the following after reading Gary Lawlor's article on the brachistochrone last March [103 (1996) 242–249]:

> The phrase "a marble rolling without friction," which appears throughout the paper, is, at best, an oxymoron. A marble can roll *only* if friction is present... if there were no friction present then the marble would simply slide down any hill without rolling—friction provides the torque that causes the marble to rotate as it translates down the hill. Furthermore, if the angle of the tangent to the hill is too steep, then the frictional force is too weak to keep the marble from slipping initially, so that in large amplitude motion of a marble in a cycloid-shaped trough, one expects to have both rolling and sliding occur, at least on the steepest parts of the trough... I suspect that what the author means by "rolling without friction" is, in fact, "rolling without slipping" or, more precisely, "rolling with no non-conservative work being done." The more realistic problem of sliding and rolling with friction goes by the name of "motion with non-holonomic constraints" in more advanced classical mechanics textbooks, and, frankly, is usually avoided by them. Frank Crawford has written an interesting article about Galileo's confusion of this issue and some at-home rolling experiments; see Rolling and Slipping Down Galileo's Inclined Plane: Rhythms of the Spheres, *Amer. J. Physics* 64 (5), May, 1996.

David Fowler's lead article in last January's issue [103 (1996) 1–17] contained three figures that illustrate the behavior of the binomial coefficient function of two real variables "x choose y." Unfortunately, the fine detail of these remarkable figures was largely lost in the printing process, though a much better version of the first figure is reproduced on the cover of the August–September issue in 1995 [102 (7) (1995)]. The author has prepared an insert sheet in which these figures are reproduced with high resolution. Postscript files in various forms (corregenda.ps = 4.2MB; .pdf = .74MB; .psZ = .79MB) are available by ftp from

ftp.maths.warwick.ac.uk/pub/papers/dhf

or at

http://www.maths.warwick.ac.uk/maths/papers/dhf.html.

For a printed copy, contact the author at University of Warwick.

Harold Boas writes that the question of differentiability of the ruler function discussed in a Note by Richard Darst and Gerald Taylor in last May's issue [103 (1996) 415–416] has a long history. Some previous MONTHLY papers that address this problem are: Gerald J. Porter, On the differentiability of a certain well-known function, 69 (1962) 142; G. A. Heuer, Functions continuous at the irrationals and discontinuous at the rationals, 72 (1965) 370–373; J. E. Nymann, An application of Diophantine approximation, 76 (1969) 668–671; Alec Norton, Continued fractions and differentiability of functions, 95 (1988) 639–643.

Roger A. Horn, *Editor*

THE AMERICAN MATHEMATICAL MONTHLY

Volume 104, Number 2 February 1997

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and notes should be sent to the Editor at the MONTHLY's Utah office:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send your email address and 3 copies of the complete manuscript (including all figures with captions and lettering), typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper.

Letters to the Editor on any topic, for publication or for private reading, are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

See the MONTHLY section of MAA Online for current information such as contents of issues, descriptive summaries of forthcoming articles, tips for authors, and preparation of manuscripts in TEX:

http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all material, typewritten on only one side of the paper.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Mr. JOE WATSON, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

Yueh-Gin Gung and Dr. Charles Y. Hu Award for Distinguished Service to Deborah Tepper Haimo

Carole B. Lacampagne

From her early days at Radcliffe, studying with Hassler Whitney, and later meeting her husband-to-be in a Harvard class taught by Saunders Mac Lane, Deborah Tepper Haimo has had a love affair with mathematics. Her dedication to the MAA began soon after Frank and Deborah Haimo's marriage when they both joined the Association. It culminated in her becoming president of the MAA in 1992. She continues to be an active and influential member of the mathematics community.

All presidents of the MAA are called upon to provide a heavy service effort. But Professor Haimo has gone beyond the normal presidential service by her reorganization of the cumbersome MAA committee structure, by her personal devotion to obtaining the recognition of outstanding teaching in each MAA Section of the country, by creating the national awards bearing the name of her late husband and herself, and by encouraging the participation of women in mathematics at every level and in the Association. These are tremendously valuable achievements, worthy of the Gung-Hu Award.

Deborah Haimo recognized well before her presidency the need for reorganization of the MAA's committee structure and chaired a committee that devised the Coordinating Council system currently being used. This new structure made order out of chaos, and it is difficult to imagine carrying on with the old structure, given the complexity of the MAA today.

Professor Haimo's dedication to excellence in teaching is clear in her own well-honed lessons and in her innovative, applications-oriented teacher enhancement program for high school teachers of mathematics. Again, with typical insight, she called attention to the fact that although the Association has always claimed to value good teaching, there was nothing in the awards structure of the organization to highlight this. Outstanding expository writing (the Chauvenet, Ford, Allendoerfer, Pólya, Hasse, and Beckenbach awards) and service (the Gung-Hu Award, Certificates for Meritorious Service) are appreciated and rewarded, but there was no particular recognition for excellence in teaching. She proceeded to stir things up (quietly, as is her style) and the Association soon established sectional awards for teaching, along with three national awards.

Just to make sure that new teaching awards are likely to continue, she initiated a significant gift to the Association. This prompted the Board of Governors to name the national awards the Deborah and Franklin Tepper Haimo Awards for Distinguished College or University Teaching of Mathematics. These are now firmly established as the prestigious awards for teaching and the list of winners is indeed distinguished.

Even prior to her presidency, Professor Haimo served the Association for many years on many important committees (the Committee on the Teaching of Undergraduate Mathematics, the 1975 Nominating Committee, the Program Committee for the 1977 Meetings in St. Louis, the Committee on the Participation of Women in Mathematics, to name a few) and as a member-at-large of the Board of Governors (1974–76). In 1986–87 she was First Vice President and in 1988–1989 she served as chair of the Search Committee for an Executive Director. She is now chair of both the Nominating Committee and the Development Committee.

As a woman mathematician, Professor Haimo has not only been a role model for female students, but has written and spoken energetically about the need to give young women the opportunity to study mathematics to the limit of their abilities and interests by making sure that the climate in mathematics departments and elsewhere, is supportive, welcoming, and encouraging to them. At the same time, she has been insistent that female students, like their male counterparts, be challenged to achieve at the very highest mathematical levels, at every step from kindergarten through graduate school.

Her undergraduate years were spent at Radcliffe College, which gave her its alumnae Recognition Award in 1993. She went on to earn a Ph.D. in classical analysis at Harvard University. Although she had no connection beforehand with Franklin and Marshall College, she received an honorary degree of Doctor of Science at a special spring convocation in 1991. After faculty appointments at Washington University (St. Louis) and Southern Illinois University at Edwardsville, she took a position at the University of Missouri–St. Louis where she served as chair of the Department for some years and eventually became Professor Emerita. She has held visiting appointments at the Technion in Israel and at the Institute for Advanced Study, Princeton, as a member in 1972–73 and later as a trustee of the Association of Members.

Always interested in educational matters, aside from research supported by federal agencies, she received numerous grants for teacher education programs at the University of Missouri. She has served on numerous national and international panels and committees: mathematician for an Agency for International Development Science Team to evaluate graduate programs at Seoul National University (1974), the ETS College Level Examinations Program Committee (1986–89), the MAA/NCTM National Selection Committee for the Presidential Awards for Excellence in Science and Mathematics Teaching (1988), to name a few.

At the same time, she has never abandoned her role as a research mathematician, having published over 45 papers in classical analysis, in particular, on generalizations of the heat equation, special functions, and harmonic analysis. She has served as an associate editor of the SIAM Journal on Mathematical Analysis and the American Mathematical Monthly. She now holds an appointment at the University of California, San Diego, where she lives right near the ocean that she dearly loves.

Always active in educational matters outside mathematics as well as within, she served as a trustee of Radcliffe College between 1975 and 1981, and she has just completed her tenure as a member of the Board of Overseers of Harvard University. She is active in the American Association for the Advancement of Science and after serving on statewide committees in Missouri and Connecticut, since settling in California, she was recently appointed by the State Board of Education to the Mathematics Framework and Criteria Committee, a major assignment. Her contributions to mathematics and to education continue at a breathtaking pace.

Professor Haimo is the mother of five very talented offspring and has nine grandchildren and one great grandchild. An active sportswoman, she is an enthusiastic tennis player. She has participated and won many medals in track events at Senior Olympics in Missouri and Illinois, and even won a gold medal for racewalking recently in San Diego.

Scheduling Conflict-free Parties for a Dating Service

Bryan L. Shader and Chanyoung Lee Shader

1. THE MORAL. We are interested in a scheduling problem that a dating service might confront. The problem is an excellent vehicle for tying together fundamental concepts in an elementary linear algebra course. The solution presented illustrates the beautiful and powerful relationship among graph theory, combinatorics, and linear algebra. Without linear algebra the problem seems intractable. Yet elementary tools from linear algebra crack the problem.

In addition to attracting students' attention, the dating service problem reinforces many of the concepts identified by the Linear Algebra Curriculum Study Group as essential in an elementary linear algebra course [3]. The concepts that arise in the problem include: the different views of matrix multiplication, block matrix multiplication, matrix factorizations, the importance of rank, the relationships between rank and the determinant, and properties of determinants. The moral is: "A little linear algebra goes a long way".

2. THE DATING SERVICE PROBLEM. A dating service has as clients n recently divorced couples, $m_1 \& w_1$, $m_2 \& w_2, \ldots,$ and $m_n \& w_n$, where m_i and w_i are the man and woman, respectively, of the ith divorced couple. Each client wants the opportunity to socialize with the other clients, but refuses to be in the same room as his or her ex-spouse. Thus, the service cannot throw just one party and must arrange a sequence of parties. To avoid embarrassing situations, if m_i and w_j are invited to the same party, then $i \neq j$. In addition, in order to provide equal opportunity for each potential new couple, the dating service would like m_i and w_j to go to exactly one common party whenever $i \neq j$. For 5 divorced couples

Party 1:	m_1,	w_2, w_3, w_4, w_5
Party 2:	m_2,	w_1, w_3, w_4, w_5
Party 3:	m_3,	w_1, w_2, w_4, w_5
Party 4:	m_4,	w_1, w_2, w_3, w_5
Party 5:	m_5,	w_1, w_2, w_3, w_4

is one possible sequence of parties, and

Party 1:	m_1, m_2,	w_3, w_5
Party 2:	m_2, m_3,	w_4, w_1
Party 3:	m_3, m_4,	w_5, w_2
Party 4:	m_4, m_5,	w_1, w_3
Party 5:	m_5, m_1,	w_2, w_4

is another.

The dating service has approached its staff mathematician with the following questions:

1. Is it possible to design such a sequence of parties?
2. If so, (for economy's sake) what is the fewest number of parties that can be held?

3. What are all the possible sequences that have the fewest number of parties?
4. How should one handle more general situations (that is, when the clients aren't n recently divorced couples but there are still some incompatible couples)?

In the next section, we describe how the staff mathematician can formulate the general dating service problem as a problem in graph theory. In section 4, we use basic linear algebra to solve the problem for n divorced couples, and to discover some surprising consequences. For example, we will see that for n divorced couples the dating service must throw at least n parties, and if $n - 1$ is a prime number then there are just two different ways that the parties can be arranged.

3. GRAPH THEORY. A *bipartite graph* $B = B(M, W, C)$ consists of disjoint finite sets M and W, and a given subset C of $M \times W$. The elements of $M \cup W$ are called *vertices* and the elements of C are called *edges*. The bipartite graph $B(M, W, C)$ can be described by a diagram. Each element of $M \cup W$ is identified with a point, and the point corresponding to $m \in M$ is joined by a line segment to the point corresponding to $w \in W$ if and only if $(m, w) \in C$. For example, if $M = \{m_1, m_2, m_3, m_4, m_5\}$, $W = \{w_1, w_2, w_3, w_4, w_5\}$, and $C = \{(m_i, w_j) | 1 \le i, j \le 5, i \ne j\}$, then the diagram of $B(M, W, C)$ is:
A *biclique* of B is an ordered pair (X, Y) where X is a subset of M, Y is a subset of W, and $X \times Y$ is a subset of C. For example, if $X = \{m_1, m_2\}$ and $Y = \{w_3, w_4, w_5\}$, then (X, Y) is a biclique of the graph illustrated in Figure 1. Two bicliques (X_1, Y_1) and (X_2, Y_2) of B are *disjoint* provided $(X_1 \times Y_1) \cap (X_2 \times Y_2) = \emptyset$. A *biclique partition* of B is a collection $(X_1, Y_1), (X_2, Y_2), \ldots, (X_t, Y_t)$ of bicliques of B such that the bicliques (X_i, Y_i) and (X_j, Y_j) are disjoint for all $i \ne j$, and

$$C = (X_1 \times Y_1) \cup \cdots \cup (X_t \cup Y_t).$$

For example,

$$\begin{array}{ll} (1) & \begin{array}{c} (\{m_1, m_2\}, \{w_3, w_5\}), \\ (\{m_2, m_3\}, \{w_4, w_1\}), \\ (\{m_3, m_4\}, \{w_5, w_2\}), \\ (\{m_4, m_5\}, \{w_1, w_3\}), \\ (\{m_5, m_1\}, \{w_2, w_4\}) \end{array} & (1) \end{array}$$

is a biclique partition of the graph in Figure 1.

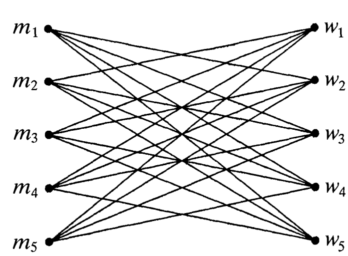


Figure 1

Since $(\{m\}, \{w\})$ is a biclique for each edge $(m, w) \in C$, the collection of all such bicliques forms a biclique partition of B. Thus, B always has at least one biclique partition. Indeed, since $\{\{m_i\}, \{w_j: (m_i, w_j) \in C\}\}$ is a biclique for each i, and $\{\{m_i: (m_i, w_j) \in C\}, \{w_j\}\}$ is a biclique for each j, B always has a biclique partition with $\min\{|M|, |W|\}$ bicliques, where $|\cdot|$ denotes the cardinality of a set. The *biclique partition number*, $\text{bp}(B)$, of B is the fewest number of bicliques among all biclique partitions of B. The preceding observation gives the upper bound

$$\text{bp}(B) \leq \min\{|M|, |W|\}. \tag{2}$$

A biclique partition of B that has exactly $\text{bp}(B)$ bicliques is called an *exact* biclique partition.

The staff mathematician can reformulate the general dating service problem in terms of bipartite graphs and biclique partitions as follows. View M as the set of gentleman clients, W as the set of lady clients, and C as the set of compatible pairs (i.e., $(m, w) \in C$ if and only if m and w are willing to attend the same party). The graph illustrated in Figure 1, is the bipartite graph in the case of 5 divorced couples. If a party is viewed as a subset of $M \cup W$, then a party with no conflicts corresponds to a biclique of B. Thus, a sequence of conflict-free parties where each compatible pair (m, w) attends exactly one common party corresponds to a biclique partition of G, and the fewest number of parties in such a sequence is the biclique partition number of B. For example, the biclique partition in (1) corresponds to the second sequence of parties described in Section 2.

Therefore, since every bipartite graph has a biclique partition, the staff mathematician knows that it is always possible to design a sequence of conflict-free parties, and that the fewest number of parties needed is the biclique partition number of the corresponding graph. To solve the problem of 5 divorced couples for the dating service, the staff mathematician needs to find the biclique partition number of the graph in Figure 1, and then classify the exact biclique partitions of this graph.

The following result (see **[1]**) determines the biclique partition number of the graph corresponding to n divorced couples and places severe restrictions on its exact biclique partitions.

Theorem 3.1. *For n an integer with $n \geq 2$, let $M = \{m_1, m_2, \dots, m_n\}$, $W = \{w_1, w_2, \dots, w_n\}$, $C = \{(m_i, w_j) \mid 1 \leq i, j \leq n, i \neq j\}$, and $B = B(M, W, C)$. Then*

$$\text{bp}(B) = n.$$

Furthermore, if $(X_1, Y_1), \dots, (X_n, Y_n)$ is an exact biclique partition of B, then there exist integers r and s such that

a) *$rs = n - 1$, $|X_i| = r$, and $|Y_j| = s$ for each $i = 1, 2, \dots, n$.*
b) *Each element of M is in exactly r of the X_i and each element of W is in exactly s of the Y_i, and*
c) *For $i \neq j$, exactly one element of $X_i \times Y_j$ is not in C.*

Theorem 3.1 shows that for n divorced couples the dating service must throw at least n parties. Moreover, for any sequence of n parties and for some integers r and s with $rs = n - 1$, we know that each party has r men and s women, each man is invited to exactly r parties, each woman is invited to exactly s parties, and for any two parties there is exactly one man who is invited to the first party and whose ex-wife is invited to the second party. The case that $n - 1$ is prime is particularly interesting, since, if $rs = n - 1$, then either $r = n - 1$, or $s = n - 1$. It follows that there are exactly two ways that the dating service can design a

sequence of n parties. Either for each man there is a party where he is the only invited man and all women (except his ex-wife) are invited, or for each woman there is a party where she is the only invited woman and all men (except her ex-husband) are invited.

In the next section, we follow **[1]** and prove Theorem 3.1 using elementary linear algebra.

4. LINEAR ALGEBRA. The problem of determining the biclique partition number of a bipartite graph can be rephrased as a matrix problem. Throughout this section, let $M = \{m_1, \dots, m_k\}$, $W = \{w_1, \dots, w_l\}$, and let B be a bipartite graph $B(M, W, C)$. The *reduced adjacency matrix* of B is the k-by-l matrix $A = [a_{ij}]$ with $a_{ij} = 1$ if $(m_i, w_j) \in C$, and $a_{ij} = 0$ otherwise. For example, the reduced adjacency matrix of the graph in Figure 1 is

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

For any subset R of M we let

$$\vec{R} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{bmatrix}$$

where $r_i = 1$ if $m_i \in R$, and $r_i = 0$ otherwise. Similarly, for any subset S of W we let

$$\vec{S} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_l \end{bmatrix}$$

where $s_i = 1$ if $w_i \in S$, and $s_i = 0$ otherwise. It is easy to verify that $\vec{R}\vec{S}^T$ is the k-by-l matrix with (i, j) entry equal to 1 if $(m_i, w_j) \in R \times S$, and equal to 0 otherwise. Thus, $\vec{R}\vec{S}^T$ is the reduced adjacency matrix of the bipartite graph

$$\big(M, W, \{(m_i, w_j) \colon m_i \in R \text{ and } w_j \in S\}\big).$$

Remark 4.0. Each entry of $\vec{R}\vec{S}^T$ is less than or equal to the corresponding entry of A if and only if (R, S) is a biclique of B.

Lemma 4.1. *Let $X_1, X_2, \dots, X_p$ be subsets of M and let $Y_1, Y_2, \dots, Y_p$ be subsets of W. Then $(X_1, Y_1), \dots, (X_p, Y_p)$ is a biclique partition of B if and only if $A = \vec{X}_1\vec{Y}_1^T + \cdots + \vec{X}_p\vec{Y}_p^T$.*

Proof: First suppose that $(X_1, Y_1), \dots, (X_p, Y_p)$ is a biclique partition of B. Let i and j be such that the (i, j) entry of A equals 1. Then there is an edge joining m_i and w_j in B, and this edge is contained in exactly one of the bicliques $(X_1, Y_1), \dots, (X_p, Y_p)$, say (X_s, Y_s). Thus the (i, j) entry of $\vec{X}_s\vec{Y}_s^T$ equals 1, and the (i, j) entry of $\vec{X}_r\vec{Y}_r^T$ equals 0 for $r \neq s$. It follows that the (i, j) entry of $\vec{X}_1\vec{Y}_1^T + \cdots + \vec{X}_p\vec{Y}_p^T$ equals 1. Now let i and j be such that the (i, j) entry of A equals 0. Then B does

not contain the edge joining m_i and w_j and there is no r such that $m_i \in X_r$ and $w_j \in Y_r$. Thus the (i, j) entry of $\vec{X}_r \vec{Y}_r^T$ equals 0 for $r = 1, 2, \ldots, p$, and hence the (i, j) entry of $\vec{X}_1 \vec{Y}_1^T + \cdots + \vec{X}_p \vec{Y}_p^T$ equals 0. Therefore, $A = \vec{X}_1 \vec{Y}_1^T + \cdots + \vec{X}_p \vec{Y}_p^T$.

Conversely suppose that

$$A = \vec{X}_1 \vec{Y}_1^T + \cdots + \vec{X}_p \vec{Y}_p^T. \tag{3}$$

Since each $\vec{X}_r \vec{Y}_r^T$ is a matrix of 0's and 1's, the entries of each $\vec{X}_r \vec{Y}_r^T$ are less than or equal to the corresponding entries of A. It follows from Remark 4.0 that each (X_r, Y_r) is a biclique of B. Consider any edge, say the edge joining m_i and w_j, of B. Since the (i, j) entry of A equals 1 and each $\vec{X}_r \vec{Y}_r^T$ is a matrix of 0's and 1's, equation (3) implies that there is a unique s such that the (i, j) entry of $\vec{X}_s \vec{Y}_s^T$ is nonzero. Hence, the edge joining m_i and w_j is contained in exactly one of the bicliques $(X_1, Y_1), \ldots, (X_p, Y_p)$. Therefore, $(X_1, Y_1), \ldots, (X_p, Y_p)$ forms a biclique partition of B. ■

Remark 4.2. Let $X_1, \ldots, X_p$ be subsets of M and let $Y_1, \ldots, Y_p$ be subsets of W. Let X be the k-by-p matrix whose ith column is $\vec{X}_i$ and let Y be the p-by-l matrix whose ith row is $\vec{Y}_i^T$. Then by block multiplication of matrices we see that

$$XY = \vec{X}_1 \vec{Y}_1^T + \vec{X}_2 \vec{Y}_2^T + \cdots + \vec{X}_p \vec{Y}_p^T.$$

Thus, a biclique partition of B into p bicliques gives rise to the factorization $A = XY$, where X is a k-by-p $(0,1)$-matrix and Y is a p-by-l $(0,1)$-matrix. Conversely, it is easy to verify that if X is a k-by-p $(0,1)$-matrix and Y is a p-by-l $(0,1)$-matrix with $A = XY$, then $(X_1, Y_1), (X_2, Y_2), \ldots, (X_p, Y_p)$ is a biclique partition of B, where $\vec{X}_i$ is the ith column of X and $\vec{Y}_i^T$ is the ith row of Y.

In particular, it follows that the biclique partition number of B is the smallest p such that $A = XY$, where X is a k-by-p $(0,1)$-matrix and Y is p-by-l $(0,1)$-matrix. Since $p \geq \operatorname{rank}(X) \geq \operatorname{rank}(XY)$, we have the lower bound

$$\operatorname{bp}(B) \geq \operatorname{rank}(A). \tag{4}$$

While it is easy to compute the rank of A, it seems quite difficult to compute $\operatorname{bp}(B)$ for arbitrary bipartite graphs B (see [**6**]).

There are bipartite graphs B for which $\operatorname{bp}(B) > \operatorname{rank}(A)$. For example, let B be the bipartite graph whose reduced adjacency matrix is

$$a = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Clearly, A has rank 3. It is easy to see that each biclique of B has at most 2 edges. Since B has 8 edges, this implies that $\operatorname{bp}(B) \geq 4$. Therefore, the biclique partition number of B is greater than the rank of its reduced adjacency matrix.

For the remainder of this section we specialize to the n divorced couples case with $n \geq 2$. Thus, $M = \{m_1, m_2, \ldots, m_n\}$, $W = \{w_1, w_2, \ldots, w_n\}$ and $C = \{(m_i, w_j): i \neq j\}$. In this case $A = J_n - I_n$ where J_n is the n-by-n matrix of all 1's and I_n is the n-by-n identity matrix. It is easy to verify that $((n-1)^{-1}J_n - I_n)A = I_n$. Hence A is invertible, so $\operatorname{rank}(A) = n$. The bounds (2) and (4) permit us to conclude that

$$\operatorname{bp}(B) = n. \tag{5}$$

The following lemma is part of the folklore of linear algebra.

Lemma 4.3. *Let X be an n-by-k matrix and let Y be a k-by-n matrix. Then*

$$\det(I_n + XY) = \det(I_k + YX).$$

Proof: Check that

$$\begin{bmatrix} I_k + YX & 0 \\ X & I_n \end{bmatrix}\begin{bmatrix} I_k & Y \\ 0 & I_n \end{bmatrix} = \begin{bmatrix} I_k & Y \\ 0 & I_n \end{bmatrix}\begin{bmatrix} I_k & 0 \\ X & I_n + XY \end{bmatrix}$$

and take determinants of both sides. Recall that the determinant of a product of square matrices is the product of the determinants of the factors, and that the determinant of a block triangular matrix whose diagonal blocks are square is the product of the determinants of the diagonal blocks. ∎

Lemma 4.4. *Let X and Y be n-by-n $(0,1)$-matrices with $XY = J_n - I_n$ and $n \geq 2$. Then $XY = YX$, and there exist positive integers r, s with $rs = n - 1$ such that $XJ = JX = rJ$ and $YJ = JY = sJ$.*

Proof: Let X_i denote the ith column of X, and let Y_i^T denote the ith row of Y. Since

$$0 = \operatorname{tr}(XY) = \operatorname{tr}(YX) = \sum_{j=0}^{n} Y_j^T X_j,$$

we conclude that

$$Y_j^T X_j = 0 \qquad \text{for} \quad j = 1, 2, \ldots, n. \tag{6}$$

Let e be the n-by-1 vector of all 1's. Then $J_n = ee^T$. For $i \neq j$, we have

$$I_n + X_i Y_i^T + X_j Y_j^T = ee^T - \sum_{l \neq i, j} X_l Y_l^T. \tag{7}$$

The right-hand side of equation (7) is the sum of $n - 1$ matrices each of rank 1. Since the rank of a sum of matrices is at most the sum of the ranks of the matrices, we conclude that the n-by-n matrix $I_n + X_i Y_i^T + X_j Y_j^T$ has rank at most $n - 1$, and hence is not invertible. Thus $0 = \det(I_n + X_i Y_i^T + X_j Y_j^T)$. Using this observation and Lemma 4.3, we conclude that

$$0 = \det\left(I_n + X_i Y_i^T + X_j Y_j^T\right) = \det\left(I_n + [X_i, X_j]\begin{bmatrix} Y_i^T \\ Y_j^T \end{bmatrix}\right)$$

$$= \det\left(I_2 + \begin{bmatrix} Y_i^T \\ Y_j^T \end{bmatrix}[X_i, X_j]\right) = \det\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & Y_i^T X_j \\ Y_j^T X_i & 0 \end{bmatrix}\right)$$

$$= 1 - \left(Y_i^T X_j\right) \cdot \left(Y_j^T X_i\right).$$

Thus $1 = Y_i^T X_j \cdot Y_j^T X_i$. Since the entries of Y_i and X_j are 0's and 1's, this implies that

$$Y_i^T X_j = 1 \text{ for } i \neq j. \tag{8}$$

Since,

$$YX = \begin{bmatrix} Y_1^T \\ \vdots \\ Y_n^T \end{bmatrix} [X_1, \ldots, X_n] = [Y_i^T X_j],$$

(6) and (8) imply that $YX = J_n - I_n = XY$. Since X commutes with Y, X, and I_n, it commutes with $YX + I_n$, which is J_n. A matrix commutes with J_n if and only if its row sums and column sums are all equal, so $XJ_n = J_n X = rJ_n$ for some nonnegative integer r. Similarly, $YJ_n = J_n Y = sJ_n$ for some nonnegative integer s. Because $rsJ_n = (XY)J_n = (J_n - I_n)J_n = (n-1)J_n$, $rs = n - 1$, and the proof is complete. ■

We now have all the necessary ingredients to prove our main result, Theorem 3.1. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_p, Y_p)$ be a biclique partition of B. By (5), $p \geq n$. Let X be the n-by-p matrix whose ith column is $\vec{X}_i$, and let Y be the p-by-n matrix whose ith row is $\vec{Y}_i^T$. From Remark 4.2, $XY = J_n - I_n$. Thus, if $p = n$, Lemma 4.4 applies and statements a), b), and c) of Theorem 3.1 follow by appropriately interpreting the entries of XJ, YJ, JX, JY, and YX.

5. EPILOGUE. We conclude with some further observations and a brief discussion of related problems.

For any pair of positive integers r and s with $rs = n - 1$, there is a (0, 1)-matrix factorization $XY = J_n - I_n$ with $XJ_n = J_n X = rJ_n$ and $YJ_n = J_n Y = sJ_n$. Namely, let $X = Z + Z^2 + \cdots + Z^r$ and $Y = I_n + Z^r + Z^{2r} + \cdots + Z^{(s-1)r}$, where Z is the n-by-n matrix with 1's in positions $(1, 2), (2, 3), \ldots, (n-1, n)$, and $(n, 1)$.

If $XY = J_n - I_n$ is a (0, 1)-matrix factorization then so is $\hat{X}\hat{Y} = J_n - I_n$ where $\hat{X} = XP$, $\hat{Y} = P^T Y$, and P is a permutation matrix. The conflict-free party scheme corresponding to that of $\hat{X}\hat{Y} = J_n - I_n$ can be obtained from that corresponding to $XY = J_n - I_n$ by permuting the order of the parties according to the permutation P. Thus, one may consider these factorizations to be equivalent. As shown in **[2]**, there are matrix factorizations of $J_n - I_n$ that are not equivalent to any of those constructed in the preceding paragraph. A complete characterization of the nonequivalent factorizations of $J_n - I_n$ (equivalently, the exact biclique partitions of the graph corresponding to n divorced couples) is not yet known.

The study of biclique partitions arose first in the context of an addressing problem in computer science **[4]**. In this setting one needs to consider graphs other than bipartite graphs.

A *biclique* B of a general graph G is a subgraph of G with the property that there exist disjoint sets X and Y of vertices of G such that the edges of B are precisely those edges joining a vertex in X and a vertex in Y. Just as in the case of bipartite graphs, the *biclique partition number* of G, bp(G), is the fewest number of bicliques of G whose edges partition the edges of G. If the vertices of G are $1, 2, \ldots, n$, then the *adjacency matrix* of G is the n-by-n matrix A whose (i, j) entry equals 1 if vertex i and vertex j are joined by an edge of G, and 0 otherwise.

The following elegant result of Graham and Pollak [4], whose proof also uses elementary linear algebra, relates the biclique partition number of a graph and the eigenvalues of its adjacency matrix.

Theorem 5.1. *Let G be a graph with adjacency matrix A. Then*

$$\mathrm{bp}(G) \geq \max\{n_+ \; n_-\},$$

where n_+ and n_- denote the number of positive and negative eigenvalues of A, respectively.

The adjacency matrix of the complete graph, K_n, is $J_n - I_n$, and the eigenvalues of this matrix are $n-1, -1, -1, \ldots, -1$. Hence by the Graham-Pollak theorem, $\mathrm{bp}(K_n) \geq n-1$. Since it is easy to construct a biclique partition of K_n with $n-1$ bicliques, $\mathrm{bp}(K_n) = n-1$. There are numerous different proofs of this fact [7, 8]. Interestingly, all currently known proofs use elementary linear algebra.

Recently Alon, Sachs, and Seymour (see section 9.12 of [5]) have proposed a problem on biclique partitions. The *chromatic number* of a graph G is the fewest number of colors needed to color the vertices of G in such a way that every pair of adjacent vertices have different colors. Clearly, the chromatic number of the complete graph K_n on n vertices is n. Thus, for the complete graph, its biclique partition number is strictly less than its chromatic number. They ask if this holds in general. Namely, is it true that if G is a graph with chromatic number k, then $\mathrm{bp}(G) < k$?

REFERENCES

1. D. de Caen and D. A. Gregory, On the decomposition of a directed graph into complete bipartite subgraphs, *Ars Combin.* **23** (1987), 139–146.
2. D. de Caen, D. A. Gregory, I. G. Hughes and D. L. Kreher, Near factors of finite groups, *Ars Combin.* **29** (1990), 53–63.
3. D. Carlson, C. R. Johnson and D. C. Lay, The Linear Algebra Curriculum Study Group Recommendations for the First Course in Linear Algebra, *College Math. J.* **24** (1993), 41–46.
4. R. L. Graham and H. O. Pollak, On the addressing problem for loop switching, *Bell Syst. Tech. J.* **50** (1971), 2495–2519.
5. T. R. Jensen and B. Toft, *Graph Coloring Problems*, Wiley, New York, 1995.
6. J. Orlin, Contentment in Graph Theory: Covering graphs with cliques, *Indag. Math.* **39** (1977), 406–424.
7. G. W. Peck, A new proof of a theorem of Graham and Pollak, *Disc. Math.* **49** (1984), 327–328.
8. H. Tverberg, On decompositions of K_n into complete bipartite graphs, *J. Graph Theory* **6** (1982), 493–494.

Department of Mathematics
University of Wyoming
Laramie, WY 82071
bshader@uwyo.edu
chan@uwyo.edu

An Infinite Set of Heron Triangles with Two Rational Medians

Ralph H. Buchholz and Randall L. Rathbun

I. INTRODUCTION. If we denote the sides of a triangle by (a, b, c) then the area is given by

$$\Delta = \sqrt{s(s-a)(s-b)(s-c)} \tag{1}$$

where $s = (a+b+c)/2$ is the *semiperimeter*. This formula is usually attributed to Heron of Alexandria circa 100 BC – 100 AD. However, it was already known to Archimedes prior to 212 BC [**5**, p. 105].

Our investigation is limited to triangles with rational sides. Even with sides of rational length, "Heron's" formula shows that the area need not be rational; any triangle with three rational sides and rational area is called a *Heron triangle*. The smallest such triangle with integer sides is the familiar (5, 4, 3) right triangle (with area 6) shown in Figure 1.

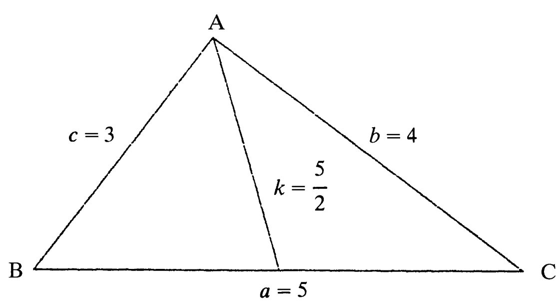


Figure 1. The (5,4,3) right angle.

If we let (k, l, m) denote the medians that are incident with the respective sides (a, b, c), they can be expressed in terms of the sides:

$$k = \frac{1}{2}\sqrt{2b^2 + 2c^2 - a^2}\,,\ l = \frac{1}{2}\sqrt{2c^2 + 2a^2 - b^2}\,,\ m = \frac{1}{2}\sqrt{2a^2 + 2b^2 - c^2}\,. \tag{2}$$

The medians of the (5, 4, 3) triangle are $(k, l, m) = \left(5/2, \sqrt{13}/2, \sqrt{73}/2\right)$. This triangle has rational area and one rational median—from the midpoint of the hypotenuse to the vertex at the right angle. It is an interesting exercise to prove that integer right triangles have precisely one rational median [1, p. 31]—the median to the hypotenuse.

But can any Heron triangle have *two* rational medians? In 1905, Schubert [**3**, p. 199] claimed that no such triangle could exist. As Dickson points out [**3**, p. 208], Schubert's proof was flawed but no such triangle was forthcoming. Despite this flaw, the parametrization used by Schubert turns out to be extremely useful in helping to uncover a key underlying pattern.

II. THE SCHUBERT PARAMETERS. Consider the triangle in Figure 2, showing one of the medians with its adjacent angles. If we apply the trigonometric identity

$$\cot\left(\frac{\alpha}{2}\right) = \frac{\sin\alpha}{1-\cos\alpha}$$

to the angle α_a say, in Figure 2, then it is clear that the corresponding half-angle cotangent is rational only if $\sin\alpha_a$ and $\cos\alpha_a$ are rational.

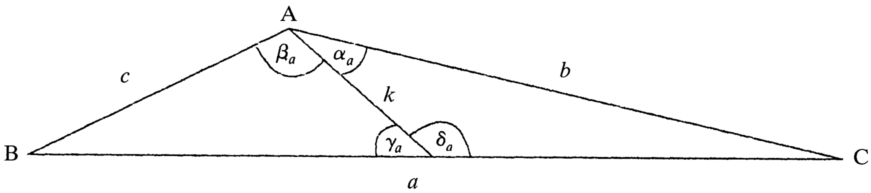


Figure 2. The angles related to Schubert's parameters.

Since

$$\sin\alpha_a = \frac{\Delta}{bk} \quad \text{and} \quad \cos\alpha_a = \frac{b^2+k^2-(a/2)^2}{2bk},$$

we see that $\sin\alpha_a$, $\cos\alpha_a$ and hence $\cot(\alpha_a/2)$ are rational for any Heron triangle with a rational median k. The same argument applies to all the angles $\alpha_a, \beta_a, \gamma_a, \delta_a$ adjacent to median k so all the half-angle cotangents are rational in this case. To ensure an unambiguous naming scheme for these parameters we impose a counter-clockwise orientation on the triangle around its centroid. Then the angles that the median to side a makes with the triangle, beginning with the two at the vertex, are labeled $\alpha_a, \beta_a, \gamma_a, \delta_a$ as in Figure 2. The respective half-angle cotangents are denoted by M_a, P_a, X_a, Y_a; see Table 1. We call the set of rational numbers (M, P, X, Y) *Schubert parameters*; it is understood that if no subscript is present then the parameters are all obtained from the same median. For the $(5,4,3)$ Heron triangle, we obtain $(M_a, P_a, X_a, Y_a) = \left(\frac{3}{1}, \frac{2}{1}, \frac{4}{3}, \frac{3}{4}\right)$.

TABLE 1. SCHUBERT PARAMETERS FOR A TRIANGLE WITH SIDES (a, b, c)

$M_a = \dfrac{4\Delta}{4bk+a^2-3b^2-c^2}$	$P_a = \dfrac{4\Delta}{4ck+a^2-b^2-3c^2}$
$X_a = \dfrac{4\Delta}{2ak-b^2+c^2}$	$Y_a = \dfrac{4\Delta}{2ak+b^2-c^2}$

The half-angle cotangents X and Y satisfy $XY = 1$, while the three half-angle cotangents M, P, and X satisfy an important relationship first proved by Schubert:

$$\left(M - \frac{1}{M}\right) - \left(P - \frac{1}{P}\right) = 2\left(X - \frac{1}{X}\right). \tag{3}$$

Although only two parameters suffice to describe any triangle, we usually consider three parameters (M, P, X). It is important to note that if (M, P, X) does satisfy equation (3), then so do 32 related 3-tuples. These occur because equation (3) is invariant under the following operations:

(i) replace any parameter by its negated inverse, or
(ii) interchange M and P while also inverting X, or
(iii) simultaneously invert all three of the parameters.

Since all such 3-tuples correspond to the same Heron triangle, we occasionally use an alternate representation.

Conversely, if we know any set of Schubert parameters, (M, P, X) say, then we can calculate the ratio of the sides (a, b, c) from

$$\frac{a}{c} = \frac{2(X + X^{-1})}{P + P^{-1}} \qquad \frac{b}{c} = \frac{M + M^{-1}}{P + P^{-1}}. \tag{4}$$

This specifies the triangle up to homothety (a similarity transformation), which is sufficient for our purposes.

In the process of trying to describe all rational-sided triangles with three rational medians the first author discovered that any rational-sided triangle, (a, b, c), with two rational medians is given by the parametrization (see [**1**, p. 38])

$$\begin{aligned} a &= \tau\{(-2\phi\theta^2 - \phi^2\theta) + (2\theta\phi - \phi^2) + \theta + 1\} \\ b &= \tau\{(\phi\theta^2 + 2\phi^2\theta) + (2\theta\phi - \theta^2) - \phi + 1\} \\ c &= \tau\{(\phi\theta^2 - \phi^2\theta) + (\theta^2 + 2\theta\phi + \phi^2) + \theta - \phi\} \end{aligned} \tag{5}$$

for (τ, ϕ, θ) constrained such that $\tau > 0$, $0 < \theta, \phi < 1$, and $\phi + 2\theta > 1$. In this case, if the parameters (τ, θ, ϕ) are rational, then the corresponding triangle must have rational sides and two rational medians, namely k and l, but not necessarily rational area. The scaling factor τ is usually set to one. Solving for θ and ϕ gives

$$\theta = \frac{c - a \pm \sqrt{2c^2 + 2a^2 - b^2}}{a + b + c} \qquad \text{and} \qquad \phi = \frac{b - c \pm \sqrt{2b^2 + 2c^2 - a^2}}{a + b + c}. \tag{6}$$

Any triangle obtained from a rational triple (M, P, X) has rational sides, rational area, and one rational median, while a triangle obtained from a rational pair (θ, ϕ) has rational sides and two rational medians. It is the unveiling of the interplay of these two parametrizations of a triangle that ultimately allows us to make progress on the question mentioned in the introduction.

III. SEARCH RESULTS AND HINT OF A CONNECTION. In 1986, both authors, unaware of each other's work, began searching for Heron triangles with two rational medians. One particularly efficient method is to enumerate over the rational parameters (θ, ϕ) in equations (5) and then check if the area of the corresponding triangle is rational. This technique allowed us to obtain the last two triangles in Table 2; meanwhile, naïve exhaustion struggled to reach the fourth triangle in the list.

So Heron triàngles with two rational medians do exist. Naturally we wondered how to find, or better yet generate, more such triangles. The first author noted that

TABLE 2. SIDES, MEDIANS, AREA OF DISCOVERED HERON TRIANGLES

Sides			Medians		Area
a	b	c	k	l	
73	51	26	$\frac{35}{2}$	$\frac{97}{2}$	420
626	875	291	572	$\frac{433}{2}$	55440
4368	1241	3673	1657	$\frac{7975}{2}$	2042040
14791	14384	11257	$\frac{21177}{2}$	11001	75698280
28779	13816	15155	$\frac{3589}{2}$	21937	23931600
1823675	185629	1930456	$\frac{2048523}{2}$	$\frac{3751059}{2}$	142334216640

the first, second, fifth, and sixth triangles of Table 2 have related internal angles and asked how this could be exploited.

IV. DISCOVERY OF THE SEQUENCE OF SQUARES. In October 1989, the second author discovered a remarkable connection between the X_a and X_b parameters of related triangles. By selecting the "appropriate" Schubert parameters and inverting where necessary (denoted by an asterisk), it became possible to arrange the four triangles into a logical chain such that the M_b parameter from one triangle was equal to the P_a parameter of the next triangle. We label these first four triangles of the chain (see Table 3) by level 1, 2, 3 and 4 respectively, and insert the degenerate triangle (2, 1, 1), with rational area and medians, at level 0 to initialise the chain.

TABLE 3. TRIANGLES WITH A COMMON $\{M_b(i), P_a(i+1)\}$ RATIO.

Level i	Triangle	$M_a(i)$	$P_a(i)$	$X_a(i)$	$M_b(i)$	$P_b(i)$	$X_b(i)$
0	(2, 1, 1)	$\frac{3}{2}$	$\frac{2}{3}$	$\frac{3}{2}$	$\frac{2}{3}$	$\frac{3}{2}$	$\frac{2}{3}$
1	1st	$\frac{4}{1}$	$\frac{2}{3}$	$\frac{8}{3}$	$\frac{35}{6}$	$\frac{84}{5}$	$\frac{7}{40}$
2	2nd	$\frac{18}{1}$	$\frac{35}{6}$	$\frac{63}{10}$	$\frac{176}{105}$*	$\frac{77}{360}$*	$\frac{99}{32}$*
3	5th	$\frac{75}{98}$	$\frac{176}{105}$	$\frac{539}{800}$	$\frac{3080}{111}$	$\frac{14504}{275}$	$\frac{147}{1850}$
4	6th	$\frac{1344}{605}$	$\frac{3080}{111}$	$\frac{363}{4736}$	$\frac{3256}{165585}$*	$\frac{5312}{255189}$*	$\frac{36480}{70301}$*

The crucial observation occurred by comparing the $X_b(i)$ and $X_a(i+1)$ ratios of consecutive triangles. From levels 1 and 2 we observed that $(40 \cdot 7)/(63 \cdot 10) = \left(\frac{2}{3}\right)^2$. Similarly, levels 2, 3 and 3, 4 imply that $(99 \cdot 32)/(800 \cdot 539) = \left(\frac{3}{35}\right)^2$ and $(147 \cdot 1850)/(363 \cdot 4736) = \left(\frac{35}{88}\right)^2$. In other words, there is a distinct pattern of rational squares in the first few products of the numerators and denominators of the $X_b(i)$ and $X_a(i+1)$ parameters. Furthermore, the denominator of one square becomes the numerator of the next square. Now all one needs to specify the next triangle in the chain is the denominator of the X product ratio since this would determine $P(i+1)$, $X(i+1)$ and hence $M(i+1)$ via Schubert's equation. For example, we set $P_a(5) = M_b(4)$. Then since

$$\frac{36480 \cdot 70301}{\text{numerator}(X_a(5)) \cdot \text{denominator}(X_a(5))} = \left(\frac{88}{k}\right)^2$$

and since $P_a(5)$ and $X_a(5)$ must lead to a rational value of $M_a(5)$ in Schubert's equation (3), one finds that $k = 37$ and hence $X_a(5) = \frac{780330}{581}$. Now calculate the Schubert parameters corresponding to the other rational median in this triangle and repeat the process. This leads to the sequence of ratios

$$\left(\frac{1}{2}\right)^2, \left(\frac{2}{3}\right)^2, \left(\frac{3}{35}\right)^2, \left(\frac{35}{88}\right)^2, \left(\frac{88}{37}\right)^2, \left(\frac{37}{4731}\right)^2, \left(\frac{4731}{107134}\right)^2, \ldots$$

This permitted us to generate the next few triangles. For example, the fifth Heron-2-median triangle has sides given by (2442655864, 2396426547, 46263061).

V. CONNECTION TO SOMOS SEQUENCES. There the matter stood for 5 years, until the two authors were able to re-establish contact. The main question was: How was the rational square sequence determined, and could a formula be found for it? After intense correspondence from late 1994 to early 1995, we obtained some interesting results.

The problem with the method described in the previous section is that it requires the factorization of numbers that are growing very rapidly. Furthermore, there is still some ambiguity about inverting certain parameters and not others.

We found that all of the (M, P, X) parameters could be formed as a combination of two series. Notice that the numerator of the X_b parameter in Table 4 is the product $a_1 \cdot a_2 \cdot a_3 \cdot a_4$ and the denominator is likewise the product $b_1 \cdot b_2 \cdot b_3 \cdot b_4$,

TABLE 4. DECOMPOSITION OF THE PARAMETER $X_b(i)$

	Numerator Factors				Denominator Factors				Parameter
i	a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4	$X_b(i)$
0	$2 \cdot 1$	-1	-1	1	$1 \cdot 1$	1	3	1	$\frac{2}{3}$
1	$1 \cdot 1$	1	1	-7	$2 \cdot 2$	2	5	-1	$\frac{7}{40}$
2	$2 \cdot 2$	1	1	8	$1 \cdot 3$	3	11	1	$\frac{32}{99}$
3	$1 \cdot 3$	-7	-7	-1	$2 \cdot 5$	5	37	1	$-\frac{147}{1850}$
4	$2 \cdot 5$	8	8	-57	$1 \cdot 11$	11	83	-7	$\frac{36480}{70301}$
5	$1 \cdot 11$	-1	-1	391	$2 \cdot 37$	37	274	8	$\frac{4301}{6001696}$
6	$2 \cdot 37$	-57	-57	-455	$1 \cdot 83$	83	1217	-1	$\frac{109393830}{8383913}$

where each of the a_i and b_i are shifts of one or another of two special sequences. There are similar relationships for all the Schubert parameters for our set of

TABLE 5. THE S AND T SERIES

i	1	2	3	4	5	6	7	8	9	10
S_i	1	1	2	3	5	11	37	83	274	1217
T_i	1	-1	1	1	-7	8	-1	-57	391	-455

triangles in terms of these two series, which we denote by S and T. We observed that each series (see Table 5) seemed to satisfy an order eight recurrence, namely,

$$S_i = \frac{2^{\chi(i)} \cdot 3^{\chi(i+1)} \cdot S_{i-7} \cdot S_{i-1} + S_{i-4}^2}{S_{i-8}} \qquad \text{and}$$

$$T_i = \frac{6^{\chi(i+1)} \cdot T_{i-7} \cdot T_{i-1} + T_{i-4}^2}{T_{i-8}},$$

$$\text{where } \chi(i) = \begin{cases} 0 \text{ if } i \text{ is even} \\ 1 \text{ if } i \text{ is odd.} \end{cases}$$

Since these two series were so fundamental, one author sent a query to the *On-Line Encyclopædia of Integer Sequences* (`sequences@research.att.com`), authored by Neil J. A. Sloane. It quickly posted back that the first, S series, was indeed a Somos 5 sequence [**4**, p. 41], and gave the recursion formula

$$A_i = \frac{A_{i-1} \cdot A_{i-4} + A_{i-2} \cdot A_{i-3}}{A_{i-5}}. \tag{7}$$

We realised that the T series satisfied the same recurrence with different initial terms. In terms of the order 5 recurrence we have

$$S_i = \begin{cases} 1,1,2,3,5 & \text{for } i = 1,\dots,5 \\ A_i & \text{for } i \geq 6, \end{cases} \qquad T_i = \begin{cases} 1,-1,1,1,-7 & \text{for } i = 1,\dots,5 \\ A_i & \text{for } i \geq 6. \end{cases} \tag{8}$$

The half-angle cotangents of our chain of Heron triangles with two rational medians are given in terms of the series S and T by

$$\begin{aligned} M_a(i) &= -\frac{S_{i+1}\cdot S_{i+2}^2\cdot T_i}{S_i\cdot T_{i+1}\cdot T_{i+2}^2} & M_b(i) &= \frac{S_{i+1}\cdot S_{i+4}\cdot T_{i+1}\cdot T_{i+4}}{S_{i+2}\cdot S_{i+3}\cdot T_{i+2}\cdot T_{i+3}} \\ P_a(i) &= -\frac{S_{i+1}\cdot S_{i+2}\cdot T_{i+1}\cdot T_{i+2}}{S_i\cdot S_{i+3}\cdot T_i\cdot T_{i+3}} & P_b(i) &= -\frac{S_{i+2}^2\cdot S_{i+3}\cdot T_{i+4}}{S_{i+4}\cdot T_{i+2}^2\cdot T_{i+3}} \\ X_a(i) &= 2^{(-1^{i+1})}\cdot\frac{S_i\cdot S_{i+2}^2\cdot T_{i+3}}{S_{i+3}\cdot T_i\cdot T_{i+2}^2} & X_b(i) &= 2^{(-1^i)}\cdot\frac{S_{i+1}\cdot T_{i+2}^2\cdot T_{i+4}}{S_{i+2}^2\cdot S_{i+4}\cdot T_{i+1}}. \end{aligned} \tag{9}$$

Equations (9) permitted us to rapidly compute many corresponding triangles using multiprecision packages (`MAPLE` and `PARI`) and each such triangle invariably had rational area and two rational medians.

VI. SEARCHING FOR A CLOSED FORM FOR S AND T SEQUENCES. Having obtained recurrence relations for S_i and T_i, we hoped that a closed formula would allow us to prove some of the results that we had so far observed only numerically. A second posting to the `sci.math.research` newsgroup prompted a number of interesting responses but by far the most impressive came from Noam Elkies, who gave two closed formulae for the S_i sequence and indirectly provided a formula for the T_i sequence. What follows borrows heavily from his reply.

Numerical evidence suggests that the sequence S_i also satisfies recurrence relations of the form

$$\begin{aligned} S_{i-2}S_{i+2} &= 2S_{i-1}S_{i+1} - S_i^2 \qquad \text{if } i \text{ is even,} \\ S_{i-2}S_{i+2} &= 3S_{i-1}S_{i+1} - S_i^2 \qquad \text{if } i \text{ is odd.} \end{aligned}$$

It is possible to combine these into a single identity by defining

$$\sigma_i = \begin{cases} S_i, & \text{if } i \text{ is even,} \\ rS_i, & \text{if } i \text{ is odd.} \end{cases}$$

Replacing S_i with σ_i or σ_i/r as appropriate and then equating the preceding two recurrences, one finds that $r = \sqrt[4]{2/3}$. Hence, the σ_i satisfy the recurrence relation

$$\sigma_{i-2}\,\sigma_{i+2} = \sqrt{6}\,\sigma_{i-1}\sigma_{i+1} - \sigma_i^{\,2}.$$

Because of the similarity of this to a Somos recurrence on sequences of elliptic theta functions, one attempts to fit a solution of the form

$$\sigma_i = bu^{i^2}\sum_{n=-\infty}^{+\infty} q^{n^2}z^{in}. \tag{10}$$

In fact, the parameters q, z, b, u can be obtained numerically from the condition that the formula for σ_{i} hold for the initial values. This leads to

$$\begin{aligned} q &= 0.022089428110979335573560 88\ldots \\ z &= 0.11419420416002380489213 21\ldots \\ b &= 0.9576898995913810138013844\ldots \\ u &= 0.7889128685374661530379575\ldots \end{aligned}$$

The theta function (10) is rapidly convergent and so we have a numerical, closed form expression to evaluate each σ_i and hence each S_i. Using the initial conditions for the T-sequence would lead to a similar theta function.

However, the numbers S_i can also be obtained "arithmetically" from the elliptic curve $\mathbb{C}^*/q^{2\mathbb{Z}}$ associated to our theta functions. By

(i) computing the j-invariant $j(E) = j(q^2)$ as a real number,
(ii) using its continued fraction to recognize $j(E)$ as the rational $11^6/612$,
(iii) computing the x-coordinate of the point z on the curve $\mathbb{C}^*/q^{2\mathbb{Z}}$, which determines the correct quadratic twist, and
(iv) reducing to standard minimal form,

Elkies finds the elliptic curve

$$E: y^2 + xy = x^3 + x^2 - 2x,$$

which is curve #102-A1 in Cremona's tables [2]. It has a point of order 2 at $(0,0)$ and an infinite order point at $P = (x, y) = (2, 2)$. For $i = 1, 2, 3, 4, \ldots$ the x-coordinate of the i-th multiple of P on E in lowest terms is

$$\frac{2\cdot 1^2}{1^2}, \frac{1^2}{1^2}, \frac{2\cdot 2^2}{1^2}, \frac{3^2}{1^2}, \frac{2\cdot 5^2}{7^2}, \frac{11^2}{8^2}, \frac{2\cdot 37^2}{1^2}, \ldots.$$

Indeed, the numerator of $i * P$ is always S_i^2 or $2S_i^2$ according as i is even or odd. Notice that the denominator is precisely T_i^2. The two sequences are very closely connected. Not only do they satisfy the same recurrence relation, but the initial conditions are no longer arbitrary; given one it is possible to construct the other.

Unfortunately, we were not able to use either of these closed forms to prove that the triangles generated from equations (9) and (4) always have rational area. However, the elliptic curve does turn up again and leads to such a proof from a different direction.

VII. TRIANGLES IN THE $\theta\phi$-PLANE LEAD TO FIVE ELLIPTIC CURVES. At this stage we used equations (9), (4), and (6) to generate the values of θ and ϕ corresponding to the first 100 terms of the two Somos sequences S_i and T_i. We plotted these parameters, considered as points corresponding to distinct Heron triangles with two rational medians, in the $\theta\phi$-plane (Figure 3) and the structure here was a surprise.

Rather than being randomly distributed in the region, the points seem to lie on five distinct curves. During this process we discovered that the points were being distributed to the five curves in a periodic way with a cycle length of 7. The points generated by the parameter set $(M_a(i), P_a(i), X_a(i))$ visited the curves in the order $\{1,2,3,4,1,2,5\}$. Similarly, the points generated by the set $(M_b(i), P_b(i), X_b(i))$ visited the curves in the order $\{2,1,4,3,2,1,5\}$. As a result, it was easy to isolate the rational coordinates of enough points on each curve to determine the corresponding equations:

$$C_1:\ 27\theta^3\phi^3 - \theta\phi(\theta-\phi)(8\theta^2 + 11\theta\phi + 8\phi^2) - 3\theta\phi(5\theta^2 - \theta\phi + 5\phi^2)$$
$$-(\theta-\phi)(\theta^2 + 40\phi + \phi^2) - (3\theta^2 - 7\theta\phi + 3\phi^2) - 3(\theta-\phi) - 1 = 0,$$

$$C_2:\ 3\theta^2\phi^2 - 2\theta\phi(\theta-\phi) - (\theta^2 + 6\theta\phi + \phi^2) + 1 = 0,$$

$$C_3:\ \theta\phi(\theta-\phi)^3 - (\theta^4 + 11\theta^3\phi + 3\theta^2\phi^2 + 11\theta\phi^3 + \phi^4)$$
$$-2(\theta^3 - \phi^3) + 10\theta\phi + 2(\theta-\phi) + 1 = 0,$$

$$C_4:\ \theta\phi(\theta-\phi) + \theta\phi + 2(\theta-\phi) - 1 = 0,$$

$$C_5:\ (\theta-1)^3\phi^2 + 2(\theta+1)(\theta^3 + 2\theta^2 - 2\theta + 1)\phi + (2\theta-1)(\theta+1)^3 = 0.$$

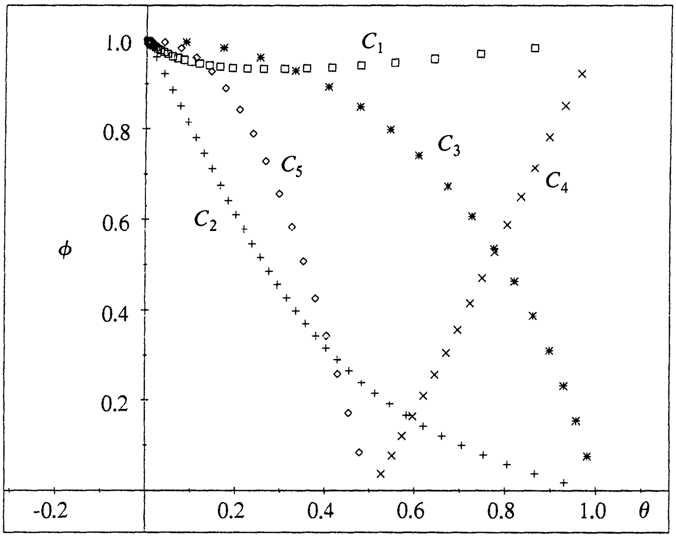


Figure 3. Heron triangles with 2 rational medians in the $\theta\phi$-plane.

We conjectured that all the rational points on these five curves produce triangles with rational area. Since the triangle has two rational medians, one can form (θ, ϕ) parameters for either median. We call these *dual* parameter sets for the triangle. The transformation that takes (θ, ϕ) to its dual point (θ', ϕ') is given by

$$\theta' = \frac{2\theta^2 + \theta\phi + \theta + \phi - 1}{3\theta\phi + \theta - \phi + 1}, \qquad \phi' = \frac{-\theta\phi - 2\phi^2 + \theta + \phi + 1}{3\theta\phi + \theta - \phi + 1}.$$

Under this mapping the curves C_1 and C_2 are dual, as are C_3 and C_4, while C_5 is self-dual. Thus it is sufficient to prove that all rational points on the curves C_2, C_4, and C_5 say, correspond to Heron triangles with two rational medians.

Next, we find that C_2, C_4 and C_5 are all birationally equivalent to the same elliptic curve so we need to prove the conjecture only for C_4, say. These three curves are quadratic in ϕ and the respective discriminants are

$$\begin{aligned} Disc\,(C_2) &= 4(4\theta^4 + 8\theta^3 + 5\theta^2 - 2\theta + 1) \\ Disc\,(C_4) &= \theta^4 + 2\theta^3 + 5\theta^2 - 8\theta + 4, \qquad \text{and} \\ Disc\,(C_5) &= 4\theta^2(\theta + 1)^2(\theta^4 + 2\theta^3 + 5\theta^2 - 8\theta + 4). \end{aligned}$$

Since we are searching for rational points on each of the curves, we require the discriminant of each to be a rational square. All the rational points that force this correspond to rational points on the elliptic curve

$$Y^2 = X^4 + 2X^3 + 5X^2 - 8X + 4.$$

For C_2, we map X to $-1/\theta$ while for C_4 and C_5 we just map X to θ. Finally we were able to prove the following

Theorem. *Every rational point on the curve*

$$C_4: \theta^2\phi - \theta\phi^2 + \theta\phi + 2\theta - 2\phi - 1 = 0$$

such that $0 < \theta, \phi < 1$ *and* $2\theta + \phi > 1$ *corresponds to a triangle with rational sides, rational area, and two rational medians.*

The proof requires several technical lemmas that will appear in a forthcoming paper. Here we just give an outline.

(i) The θ, ϕ inequalities are obtained from the triangle inequalities.
(ii) Reduce the squarefree part of the square of the area from degree 11 to degree 8 by applying the curve C_4 to Heron's formula (1).
(iii) Transform the curve C_4 to minimal Weierstraß form to obtain E, the elliptic curve found by Elkies in Section VI.
(iv) Finally, use induction in the group $E(\mathbb{Q})$ to show that any point that corresponds to a triangle with rational area leads, in all possible ways, to another point corresponding to a triangle with rational area.

VIII. TWO ISOLATED TRIANGLES. The story does not end here since two of the triangles found by computational search (the third and fourth entries of Table 2) do not lie on any of our five elliptic curves. Although these two triangles were found using equations (5), they are probably not parametrizable by equations (9) since the five curves were numerically obtained from the latter. Each of these isolated triangles has associated with it six triangles that have a rational median and rational area and share a common Schubert parameter ratio. What role these ratios play is as yet undetermined.

We are continuing further research into these two triangles, as we conjecture that all Heron triangles with two rational medians are produced by formulae similar to those we have presented in this paper. However, finding more examples like these two appears difficult.

ACKNOWLEDGMENTS. The authors greatly appreciate the assistance received from many sources, not the least of whom are those who answered our news postings. Thanks to Noam Elkies for providing the closed formulae for the Somos sequence; Ian O. Smart for correcting the recurrence relation for the σ's; Michael J. Smith for the insight to absorb those annoying negative signs into the initial conditions for the T sequence; and of course Richard K. Guy for making us aware of this problem in the first place and helping us to make each others acquaintance. We should also note that some of the triangles in Table 2 were discovered independently by Arnfried Kemnitz at about the same time.

REFERENCES

1. Buchholz, Ralph H. *On Triangles with Rational Altitudes, Angle Bisectors or Medians*, Doctoral Dissertation, Newcastle University, Newcastle, 1989.
2. Cremona, John E. *Algorithms for Modular Elliptic Curves*, Cambridge University Press, Cambridge, 1992.
3. Dickson, Leonard E. *History of the Theory of Numbers*, volume 2, Chelsea, 1952.
4. Gale, David *Mathematical Intelligencer*, volume 13, number 1, Springer, New York, 1991.
5. Kline, Morris *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, New York, 1972.

Ralph H. Buchholz
Department of Defence
Locked Bag 5076
Kingston A.C.T. 2604
AUSTRALIA
ralph@defcen.gov.au

Randall L. Rathbun
403 Marcos St Apt C
San Marcos, CA 92069-1509
USA
randall_rathbun@rc.trw.com

Connected Sprouts

T. K. Lam

The game of Sprouts is an interesting game that has been discussed by several authors (see [**1, 2, 3**]). This two-person game begins with m points on a piece of paper. A play or a move is drawing an arc joining two points or a point to itself, and then adding a new point on the arc subject to the following conditions:

(i) the arc does not cross another arc, and
(ii) the degree of each point does not exceed 3.

The game ends when no move can be made and we obtain a graph G_0 with maximum degree 3 and minimum degree 2. Call G_0 the *end graph* of the Sprouts game. It is known that the number of plays in a game is between $2m$ and $3m - 1$ inclusive [**1, 2, 3**]. If we ignore all vertices of degree 2 in G_0, then we obtain a graph G whose vertices are all of degree 3, which we call the *cubic graph* of the game.

In [**2**], Mark Copper described some properties of cubic graphs. He asked for tight lower bounds on the number of plays when the cubic graph is connected and when it is 2-connected. We answer these questions and point out an error in [**2**].

The connected graphs that can be obtained with $2m$ moves and $m \leq 3$ are given in [**1**, p. 566] and are shown in Figure 1. The original points are represented by the big dots. The small dots represent points added during a play.

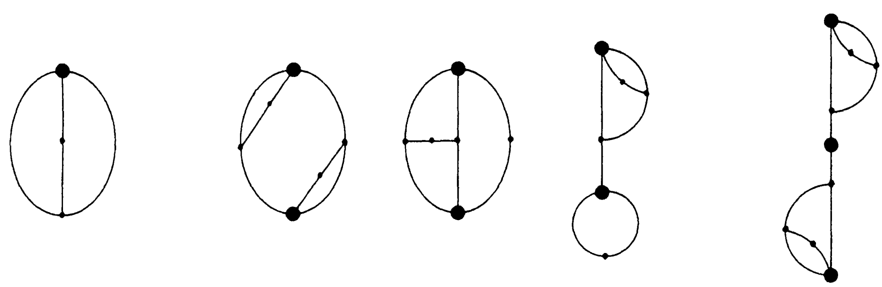


Figure 1

In [**2**], Copper stated

Proposition 5. *Suppose that the cubic graph G arises from a complete game of Sprouts on m vertices in p plays. If G is connected and $m > 2$, then $p > 2m$.*

But this is not correct, and Figure 1 gives a counterexample for $m = 3$. A correct result is:

Proposition. *Suppose that the cubic graph G arises from a complete game of Sprouts on m vertices in p plays. If G is connected, then*

$$p \geq 2m \quad \text{if } 1 \leq m \leq 3, \quad \text{and}$$
$$p \geq 2m + 1 \quad \text{if } m > 3.$$

Equality holds for some G for every value of $m = 1, 2, 3, \ldots$.

Proof: It is obvious that G is connected if and only if G_0 is connected, so we can look at G_0 instead. For $m = 1, 2$ and 3, $2m$ is just the smallest possible number of moves for a Sprouts game. Figure 1 shows that it is possible to get a connected graph in just $2m$ moves. This leaves us with the case where $m > 3$.

We will modify the proof given by Copper. Let us assume that the game of Sprouts produces a connected end graph G_0 in $2m$ moves. A quick check shows that G_0 has $3m$ vertices, $4m$ edges, and after applying Euler's formula, $m + 2$ faces. Denote the number of degree 2 vertices of G_0 by r. By counting the degrees of all the vertices, we find that

$$2r + 3(3m - r) = 2(4m), \text{ and } r = m.$$

If deleting a degree 2 vertex disconnects G_0, call this vertex a *bridge*. Note that a degree 2 vertex is a bridge if and only if it borders exactly 1 face. So, if we let b denote the number of bridges, counting the number of faces gives

$$b + 2(r - b) \leq m + 2, \text{ and } b \geq m - 2.$$

When we remove all the bridges, G_0 breaks up into $b + 1$ connected components. At least 2 of the connected components are non-trivial subgraphs to which only 1 bridge is attached. Let us call these *end-components*. Each end-component must have at least 2 interior faces. By counting the number of faces again, we find that

$$m + 2 \geq b + 4, \text{ and } b \leq m - 2.$$

Hence $b = m - 2$. Since $b + 2(r - b) = m + 2$, each face in G_0 must have a degree 2 vertex in its border. Since the interior faces of an end-component do not border a bridge, we are forced to conclude that there are exactly 2 end-components, each with 2 interior faces sharing a degree 2 vertex in their common border. We now turn to counting the number of edges. The 2 end-components must have at least 5 edges each. Each bridge has two adjacent edges and the other connected components have at least 2 edges each. This gives $2b + 2(b - 1) + 2 \times 5 = 4m$. From this equality, we know the structure of G_0. There are exactly 2 end-components with exactly 5 edges each. The other connected components have exactly 2 edges each. However, it can be verified that, for $b \geq 2$, this graph cannot be obtained from a Sprouts game. Hence, $p \geq 2m + 1$. Figure 2 illustrates such a graph when $m = 5$.

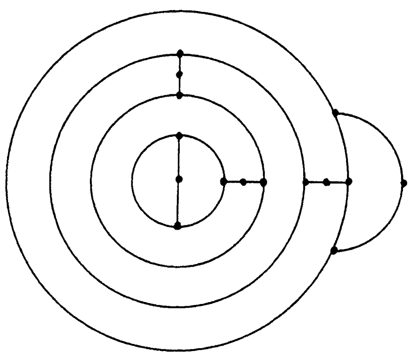


Figure 2

To show that equality is possible for each value of m, it suffices to construct a Sprouts game that ends in a connected graph in $2m + 1$ moves. We illustrate the construction with an example for $m = 5$. We first play a game of Sprouts on 2 points and get the graph shown in the middle of Figure 1. For the other points, we draw an arc connecting the point to itself in concentric circles around the first 2 points. This gives the graph shown in Figure 3a. Then we connect the circles to each other by making the moves shown in Figure 3b. This construction can be performed for every $m > 3$. ■

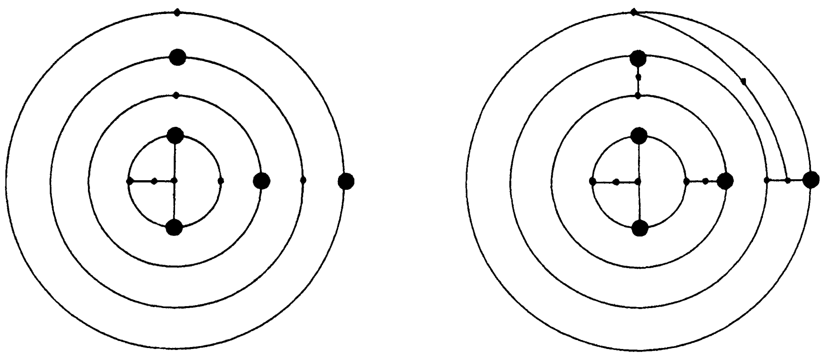


Figure 3a **Figure 3b**

Figure 3

Our proposition is equivalent to the Fundamental Theorem of Zeroth Order Moribundity stated in [**1**, p. 564].

Copper obtained a lower bound on the number of moves if the final graph obtained is 2-connected.

Theorem [2, Proposition 4]. *Suppose that the graph G arises from a complete game of Sprouts on m vertices in p plays. If G is 2-connected, then*

$$p \geq \tfrac{7}{3}m - \tfrac{2}{3}.$$

We show that the improved lower bound $\lceil (7m - 2)/3 \rceil$ is tight. Without loss of generality, we may take the m given points as lying on a circle. Joining adjacent pairs of points by arcs (and adding a new one on each arc) gives us a cycle of $2m$ points. Now take 3 consecutive points on the cycle, connect the 2 end points and then join the middle point to the newly created point. We repeat this for the next 3 consecutive points on the cycle until we are left with fewer than 3 points on the cycle. If there is no point or 1 point left, the game ends. If there are 2 points left, we make a play on them and the game ends.

Figure 4 illustrates an example for $m = 4$.

It is a simple exercise now to count the number of moves. We obtain:

$$p = \begin{cases} 7k & \text{if } m = 3k \\ 7k + 2 & \text{if } m = 3k + 1, \text{and} \\ 7k + 4 & \text{if } m = 3k + 2. \end{cases}$$

This agrees with $\lceil (7m - 2)/3 \rceil$.

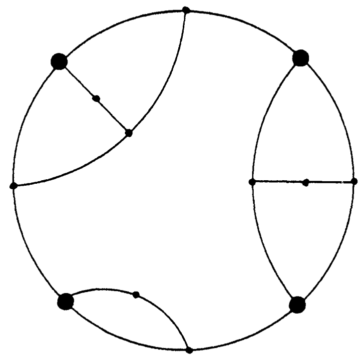


Figure 4

At the end of his article, Copper also asked which planar cubic graph are cubic graphs of some Sprouts game. Though it is tempting to conjecture that all simple planar cubic graphs can be so obtained, it is not true. A counterexample is shown in Figure 5.

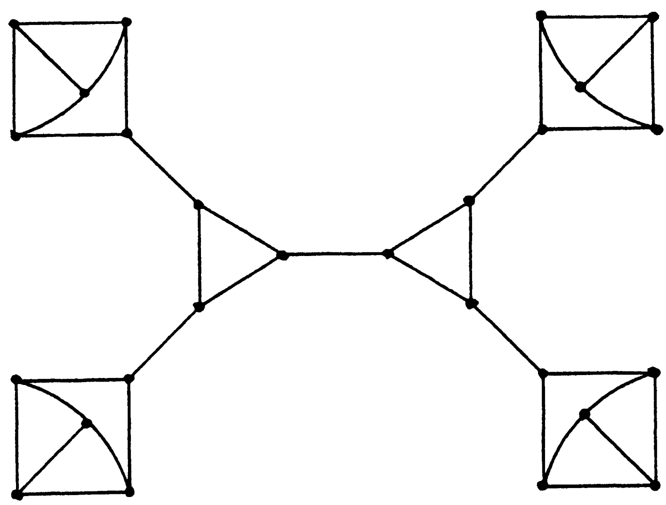


Figure 5

REFERENCES

1. E. R. Berkelemp, J. H. Conway, R. K. Guy. *Winning Ways for your Mathematical Plays, vol.* 2, Academic Press, New York, 1982.
2. Mark Copper. Graph Theory and the Game of Sprouts, *American Mathematical Monthly*, 100 (1993) 478–482.
3. Martin Gardner. *Mathematical Carnival*, Alfred A. Knopf, New York, 1975.

Department of Mathematics
National University of Singapore
Kent Ridge, S119260
Republic of Singapore
matlamtk@math.nus.sg

Regular Expressions for Program Computations

Ronald E. Prather

In the brief history of computer science, no single article has attracted more attention nor stirred more controversy than a seemingly innocuous 1968 letter to the editor (of the Communications of the ACM) by Edsger Dijkstra [**3**], titled "Goto Statement Considered Harmful." We were just emerging from the Fortran era, when programming had not yet been thought of as an art or even a craft. A diagram of program flowlines, what came to be called a *flowchart*, would very nearly resemble a plate of spaghetti—go to here and do this, then go to there and do that. Programs were "held together with baling wire," little thought being given to their organizational structure. Dijkstra argued that an "unbridled use of the goto statement" was the heart of the problem. In his opening sentence, he asserts that "the quality of programmers is a decreasing function of the density of goto statements in the programs they produce." He would later argue for the elimination of the goto statement altogether. At the time, nothing could have been more controversial. Many programmers argued that some algorithms were impossible (or at the very least, unnatural) to implement without the use of the goto. But new programming languages were on the horizon (e.g., Pascal particularly), offering such structural constituents as the *while do*, *repeat until*, *if then else*, and *case* statements, whereby one could organize a program into successive levels of refinement, giving some credence to the Dijkstra claim.

Proponents of the Dijkstra *discipline of programming* began to use a cryptic (at the time) and somewhat obscure result of Corrado Böhm and Giuseppe Jacopini [**1**] as justification for their optimism and enthusiasm. In effect, their work seemed to indicate that *any* algorithm could be written as a *structured program* [**2**], i.e., as a program involving only *repetition* (whether of the "while do" or the "repeat until" variety), *selection* (typically characterized by the "if then else" construct) and *sequence* of (possibly compound) statements, these three constructs perhaps being nested one within the other to an arbitrary depth, down to the level of the *elementary processes* of assignment, input, and output. And most importantly, there was seen to be no need whatsoever of employing the "harmful" *goto* statement.

But the Böhm-Jacopini result was not all that clearly understood at the time. Even Dijkstra himself stated only that "they *seem* to have proved the logical superfluousness of the goto statement." For it happens that it is not all that easy to give a totally convincing proof in elementary terms. We take up just such a challenge in this survey. We use the opportunity to introduce the reader to a modern software engineering framework for the investigation, using topics that are important to a contemporary computer science research area known as *software metrics*, and we further employ an algebra of regular expressions, as most commonly encountered in the *theory of automata*, all toward offering a modified form of the now-classical Böhm-Jacopini result, and thus achieving a presentation blending the new with the old.

PROGRAM FLOWGRAPHS. To begin, one must propose a model of program computation. We suppose that the reader is familiar with the notion of an *alphabet* Σ, i.e., a finite set of "symbols", and with the idea that a finite sequence of such symbols is said to constitute a *word* [**9**]. The *null word* is given a special representation ε considered as an empty sequence of symbols. It is the identity element in a monoid with respect to a binary concatenation operation on words. A collection L of words is then called a *language* (over the alphabet Σ), and we will have occasion to make use of the following operations on languages:

(1) *concatenation*; $LM = \{xy\colon x \in L \text{ and } y \in M\}$
(2) *substitution*: $L(\sigma{:}M) = \{x(\sigma{:}y){:}x(\sigma) \in L \text{ and } y \in M\}$
(3) *join*: $L \cup M = \{x\colon x \in L \text{ or } x \in M\}$

where in (2), we merely substitute individual words of M for each occurrence of the specific symbol $\sigma \in \Sigma$ in words of L.

Example 1. Suppose $\Sigma = \{a, b, c, d, e\}$ and we are given the language

$$L = \{e, ece, ecece, \dots\}$$

wherein we designate $\sigma = e$ in (2). If $M = \{a, bd\}$ then

$$L(e{:}M) = \{a, bd, aca, acbd, bdca, bdcbd, acaca, \dots\}$$

In the program model to be proposed, the symbols $\sigma \in \Sigma$ play the role of the individual elementary processes of a particular algorithm, whatever they may be. We thus are led to an abstraction of the notion of a flowchart, so as to stay within the realm of arbitrary (perhaps wildly unstructured) programs. As indicated, we do not give any attention to the specific nature of the elementary processes of an algorithm, identifying them only as symbols. Neither do we give any specific attention to the exact nature of the program decisions; they will be identified only as nodes of a flowgraph. With all of this in mind, we define a *program flowgraph F of order n over the alphabet Σ* to consist of $3n + 1$ edges among $2n + 1$ nodes or vertices, one *distinguished* and named X_F, the others evenly divided between *decision* and *junction* nodes, with an orientation of the edges such that:

(i) decision nodes have *indegree* = 1 and *outdegree* = 2;
(ii) junction nodes have *indegree* = 2 and *outdegree* = 1;
(iii) the distinguished node has *indegree* = 1 and *outdegree* = 1;
(iv) every vertex lies on a circuit through X_F (and such a circuit represents a *computation* of F).

Furthermore, each of the edges is to be labeled with a word over the alphabet Σ, the intent being to represent the sequence of elementary processes performed in traversing the individual edges of the flowgraph. The distinguished vertex X_F serves to identify a further pair of nodes, *start* = 0_F and *stop* = 1_F, owing to the unique pair of edges $1 \to X \to 0$. We omit subscripts on $X, 0, 1$ when the flowgraph F is clear from the context. Except for the vertex X, we have an underlying *cubic graph* [**6**], i.e., every vertex (except X) has degree three.

In our flowchart examples, e.g., Figure 1, decision nodes are drawn as black or solid circles, whereas junction nodes are drawn as white or open circles, merely to call attention to the distinction. The distinguished vertex X is drawn as an encircled cross (for 'X'). Because of the uniqueness of X and the property $1 \to X \to 0$, it is not necessary to identify 0 and 1 in our drawings.

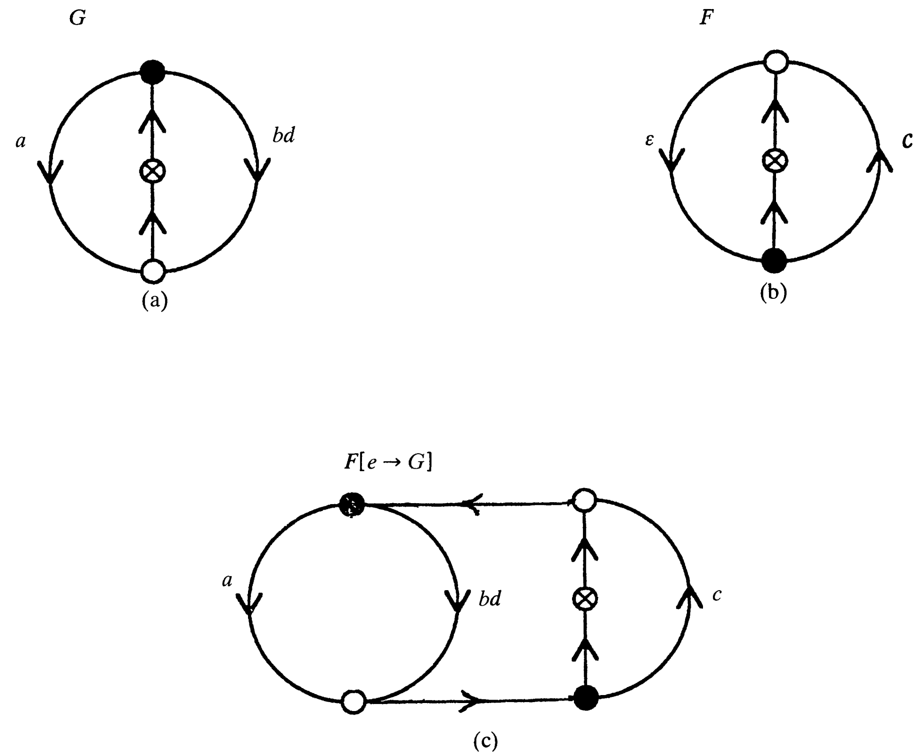


Figure 1

Example 2. While a program flowgraph can represent any program whatsoever [**11**], we will be especially interested in the so-called *structured flowgraphs*, which are built from only the sequence, selection, and repetition constructs. A sequence of elementary processes is represented as a word on the alphabet Σ, as in the case of *bd* in Figure 1(a); descriptions of more general sequences follow. A selection from among two alternating processes involves a decision, as in the case of the black circle of Figure 1(a), which by the convention we have introduced, serves as the *start* node for the flowgraph G. Thus, as computations of the selection that Figure 1(a) represents, we perform the elementary processes b and d sequentially, or we perform the elementary process a, depending on the outcome of the decision. It is important to remember that the exact nature of the decision is of no concern in the theory we are developing here. In Figure 1(b), flowgraph F is an instance of a repetition. Since the *start* node here is a junction node, and we are led (in performing a null process, symbolized by the null word) to a decision node, we may then be led to *stop* or to perform the elementary process c, depending on the outcome. In the case of the latter, we are back where we started, so that c may be executed arbitrarily many (perhaps zero) times in repetition. It is well that such computations first be understood in the context of small examples, as illustrated here.

Among the ways for building larger flowgraphs from given ones, we will be especially interested in the following two constructions:

sequence $F \circ G$: Define $0_{F \circ G} = 0_F$ and $1_{F \circ G} = 1_G$ while merging $X_F = X_G = X_{F \circ G}$. Then replace the pair of edges $1_F \to X_F$ and $X_G \to 0_G$ by $1_F \to 0_G$ while concatenating the two labels into one.

nesting $F[e \to G]$: Assume that the edge $e = (V, W)$ of F is labeled by ε. Define $0_{F[e \to G]} = 0_F$ and $1_{F[e \to G]} = 1_F$ and also $X_{F[e \to G]} = X_F$, while (eventually) discarding X_G and the edges $X_G \to 0_G$ and $1_G \to X_G$, and forming edges $V \to 0_G$ and $1_G \to W$ in place of e, respectively.

Each is an associative operation, so that compound sequences and multiple nestings can be introduced without regard to the order of their construction.

Example 3. If G is the flowgraph of Figure 1(a), seen to represent the program fragment

if ... **then** a **else** bd,

and if F is the flowgraph of Figure 1(b), taken to represent

while ... **do** c,

(we do not identify the decisions (...)), then we may nest G into F at the edge e labeled by ε in Figure 1(b), to obtain the composite flowgraph $F[e \to G]$ shown in Figure 1(c).

Now let F be any program flowgraph over the alphabet Σ. As a way of describing the multitude of combinations of computations that might be performed in *executing* the program that F represents, we introduce the *computation set* of F, defined to be the language

$$L(F) = \{x_1 x_2 \cdots x_r \text{ such that } x_i \text{ labels } e_i\ (1 \le i \le r) \text{ and } e_1 e_2 \cdots e_r \text{ is a path from } X \text{ to } X \text{ in} F\}.$$

In effect, we thereby provide an elementary *operational semantics* for our program flowgraphs, saying that $L(F)$ is the "meaning" of F, describing as it does all possible sequences of elementary processes that could result. We note, however, that this is truly an "elementary" semantics, far less detailed than the *denotational semantics* that one ordinarily introduces in a programming language context **[14]**. But we feel that it is sufficient for our purposes.

The connection with the operations on languages, as previously introduced, is quite apparent and straightforward, as seen in the following pair of elementary rules:

$$(1)\ \ L(F \circ G) = L(F)L(G)$$
$$(2)\ \ L(F[e \to G]) = L(F)(e{:}L(G)).$$

In the latter, we view F (on the right) as a program flowgraph over the extended alphabet $\Sigma \cup \{e\}$, having introduced the new label e for the edge (e) previously labeled by ε.

Example 4. Let F and G be as given in Figure 1. Then in viewing e as a member of the extended alphabet $\{a, b, c, d\} \cup \{e\}$, we have

$$L(F) = \{e, ece, ecece, \ldots\} = L \quad \text{(Example 1)}$$
$$L(G) = \{a, bd\} = M \quad \text{(Example 1)}$$

and

$$L(F)(e{:}L(G)) = L(e{:}M) = \{a, bd, aca, acbd, bdca, bdcbd, acaca, \ldots\} = L(F[e \to G])$$

REGULAR EXPRESSION ALGEBRA. The so-called *regular expressions* [13] have found a number of applications in computer science. But mainly they are known for representing the languages that are recognized by finite state automata [7]. Our effort to confirm a version of the Böhm-Jacopini statement rests on showing that regular expressions can be seen to represent the computation set of any program flowgraph.

The *algebra of regular expressions* over the alphabet Σ is the smallest collection of expressions containing the symbols $\sigma \in \Sigma$ the distinguished elements 0, 1—and closed under the three binary operations:

(i) *sum*: $\alpha + \beta$
(ii) *product*: $\alpha\beta$
(iii) *star*: $\alpha * \beta$

while satisfying the axioms:

(1a) $\alpha + \beta = \beta + \alpha$	(1b) $\alpha + \alpha = \alpha$
(2a) $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma$	(2b) $\alpha(\beta\gamma) = (\alpha\beta)\gamma$
(3a) $\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma$	(3b) $(\alpha + \beta)\gamma = \alpha\gamma + \beta\gamma$
(4a) $0\alpha = 0$	(4b) $0 + \alpha = \alpha$
(5a) $1\alpha = \alpha$	(5b) $1 * \alpha = 1 * (1 + \alpha)$
(6a) $\alpha\beta * \gamma = \alpha(\beta * \gamma\alpha)$	(6b) $\alpha * \beta = (\alpha * \beta)\beta\alpha + \alpha$

Salomaa has shown [12] that we have here a complete (and consistent) axiomatic system, relative to the interpretation that is about to be given.

In this framework, we now present the canonical inductive interpretation of regular expressions. For each regular expression α over the alphabet Σ, we define the *language* $\lambda(\alpha)$ *represented by* α as follows. First we set

$$\lambda(\sigma) = \{\sigma\} \text{ for } \sigma \in \Sigma, \text{ and}$$
$$\lambda(0) = \varnothing \text{ and } \lambda(1) = \{\varepsilon\},$$

then inductively, we define

(i) $\lambda(\alpha + \beta) = \lambda(\alpha) \cup \lambda(\beta)$
(ii) $\lambda(\alpha\beta) = \lambda(\alpha)\lambda(\beta)$
(iii) $\lambda(\alpha * \beta) = \bigcup_{k=0}^{\infty} \lambda(\alpha)(\lambda(\beta)\lambda(\alpha))^k$

One checks that if we define *equality* of regular expressions according to the agreement

$$\alpha = \beta \text{ if and only if } \lambda(\alpha) = \lambda(\beta),$$

then the entire axiom scheme outlined above is satisfied.

We will eventually draw an explicit connection between the languages (λ) corresponding to regular expressions and the languages (L) corresponding to program flowgraphs. For the present, we conclude this brief survey of the algebra of regular expressions with an elementary result making use of substitution and relating to (2) in our earlier discussion of languages:

Lemma. *Let $\rho = \rho(\sigma)$ be a regular expression involving (among others) the symbol $\sigma \in \Sigma$. If we substitute for σ, wherever it occurs, the regular expression μ, then we obtain a composite regular expression $\rho(\sigma(\mu))$ and*

$$\lambda(\rho(\sigma(\mu))) = \lambda(\rho)(\sigma : \lambda(\mu))$$

Example 5. In relation to all four of our previous examples, if we are given the pair of regular expressions:

$$\rho = e * c \qquad \lambda(\rho) = \bigcup_{k=0}^{\infty} e(ce)^k$$

$$\mu = a + bd \qquad \lambda(\mu) = \{a, bd\}$$

then

$$\begin{aligned} \rho(e(\mu)) &= (a + bd) * c \\ \lambda(\rho(e(\mu))) &= \lambda((a + bd) * c) \\ &= \bigcup_{k=0}^{\infty} \lambda(a + bd)(\lambda(c)\lambda(a + bd))^k \\ &= \bigcup_{k=0}^{\infty} \{a, bd\}(c\{a, bd\})^k \\ &= \lambda(\rho)(e:\lambda(\mu)) \end{aligned}$$

FLOWGRAPH DECOMPOSITION. The field of "software metrics" [4] falls within the general domain of software engineering. Whereas *software engineering* has come to refer to activities concerned with turning software design, development, and maintenance into a disciplined engineering practice, *software metrics* has come to be understood as just about anything within software engineering that has a quantifiable feel to it. Most broadly interpreted, this would include anything having to do with predicting software product costs, measuring and improving programming productivity, and measuring and predicting the quality and complexity of software products. The majority of its practitioners, however, intend a somewhat narrower interpretation, whereby one seeks first of all to provide a mathematical model for the notion of a program, then over that model, to introduce various numerical-valued functions (metrics) that attempt to measure one or more of the attributes implicit in the above listing, e.g., programming cost, measures of testing complexity, projected maintenance costs, etc.

We have already introduced an appropriate flowgraph model for the software metric activity. In the important subdiscipline of "hierarchical software metrics" [11], it must be possible to evaluate a metric as a recursive operation over a certain "flowgraph decomposition," one that we are about to describe. That being done, rather than to venture off into the applications in the hierarchical metric theory, however, we use the decomposition as a vehicle for deriving a modified form of the Böhm-Jacopini result.

Recalling the flowgraph sequence and nesting constructions described earlier, our decomposition theorem will be seen to answer the "inverse question": Given an arbitrary flowgraph, how may it be decomposed using these two operations? In this connection, it will then be important to give special attention to the *prime flowgraphs*, i.e., those that are irreducible or indecomposable, with respect to both sequence and nesting. And with this in mind, the fundamental result is a recursive process, discovered and attributed anew to various researchers, that we choose to identify here as follows:

Theorem 1 (Prather-Guilieri [10]). *Every program flowgraph F has a unique decomposition*:

$$F = P_1 \circ \cdots \circ P_i\left[e_{j_i} \to F_{ij}\right] \circ \cdots \circ P_k$$

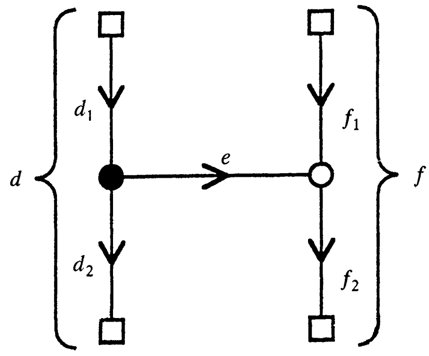


Figure 2

into a sequence of primes. Each prime P_i may have an edge e_{j_i} onto which a maximal flowgraph F_{ij} is nested. The decomposition applies (that is to say, each F_{ij} is in turn a sequence of primes, etc.), recursively, all the way down to the level of elementary processes.

Corollary. *The computation set of any program flowgraph F may be written as a concatenation:*

$$L(F) = \prod_{i=1}^{k} L\big(P_i[e_{j_i} \to F_{ij}]\big)$$

of the computation sets of its top-level primes.

Here, we have used Rule (1), where we remind the reader that the notation $P[e \to F]$ refers to a nesting of F on P. It has been shown **[11]** that the prime flowgraphs are the triple-connected ones, i.e., those for which one must cut through at least three flowlines in order to separate the flowgraph. Examples will be given shortly. But first, we turn our attention to a companion result that provides a recursive enumeration of the entire class of prime flowgraphs:

Theorem 2 (Fenton-Whitty **[5]**). *Every prime of order $n + 1$ is obtained by "grafting" (see Figure 2 where five edges—$\{d_1, d_2, e, f_1, f_2\}$ replace two—$\{d, f\}$) a decision-to-junction edge onto a prime of order n, and as a result, the prime flowgraphs can be effectively enumerated (albeit with duplication) as an infinite union*

$$\mathbf{S} = \bigcup_{n=0}^{\infty} \mathbf{S}_n,$$

where we identify (see Figure 3 and the discussion to follow):

$$\mathbf{S}_0 = \{C\}$$

$$\mathbf{S}_1 = \{D_1, D_2\}$$

$$\mathbf{S}_2 = \{E_1, E_2, E_3, E_4, E_5, E_6\}$$

etc., and in general, S_n is the subclass of primes of order n.

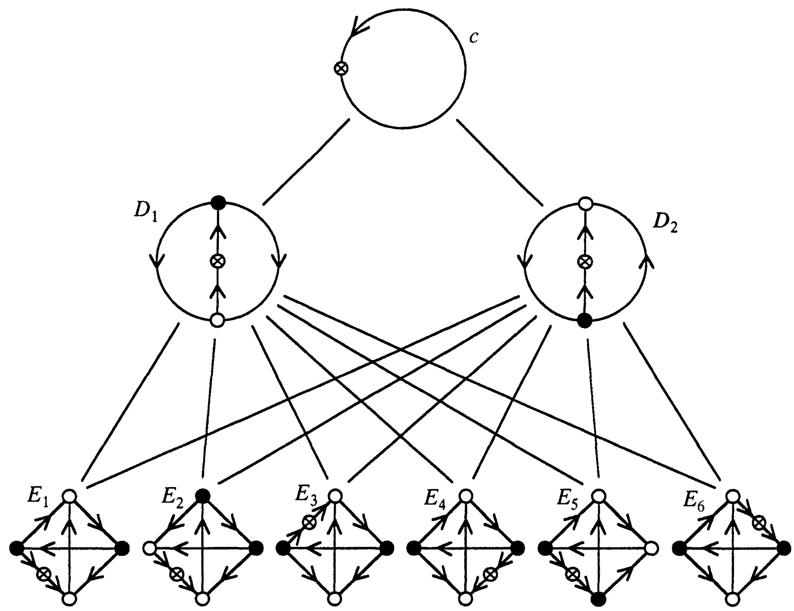


Figure 3

This result is an adaptation of a well-known principle in the theory of triply-connected cubic graphs [**6**].

The degenerate *circuit* flowgraph C (of order 0) can be used, say with its only edge labeled by a word x over the alphabet Σ, as a means of nesting a sequence of elementary processes onto a given program flowgraph. The so-called *Dijkstra* flowgraphs D_1 and D_2 can be seen to represent the programming language "if then else" and "repeat while do" constructs, respectively, the latter a combination of the conventional "repeat until" (testing for terminating the repetition *after* execution of a process) and "while do" (testing *before* execution). The reader should consult Example 4 and Example 5 for a discussion of the regular expressions that relate to the Dijkstra flowgraphs, in anticipation of the Corollary to follow.

But first, we note that the primes, in general, are really just graphs, rather than program flowgraphs. That is to say, they do not have labels on their edges, unless the edges are viewed as being labeled by themselves—as symbols of a kind of auxiliary alphabet. And yet, regular expressions can be formed with these *edges-as-labels*, leading eventually to a computational set, when each individual edge label is substituted by a genuine regular expression (over the alphabet Σ) relating to a nesting. This is the sense of the next result.

Corollary. *If $P[e_j]$ is a prime flowgraph, then there is a regular expression $\rho_P(e_j)$ over the alphabet of edges-as-symbols, such that*

$$L(P) = \lambda(\rho_P).$$

Proof: We proceed by induction over the flowgraph order n. If $n = 0$ we have the degenerate case $P = C$ that is easily handled. When $n = 1$ we have $P = D_1[a, b]$ or $P = D_2[a, b]$, where without loss of generality, we identify only the edges not

involving the distinguished node X, i.e., a for the left edge and b for the right edge in each instance of Figure 1. Clearly the regular expressions

$$\rho_{D_1} = a + b \qquad \textit{and} \qquad \rho_{D_2} = a * b$$

satisfy the indicated property, as seen in Example 2. In arguing inductively, however, it is better to show, more generally, that for each pair of vertices $u, v \neq X$, there is a regular expression ρ_{uv} representing all paths from u to v, for then we merely set $\rho_P = \rho_{X0}\, \rho_{01}\, \rho_{1X}$. Obviously this more general assertion is true for $n = 0, 1$. And if it is true for all prime flowgraphs of order n, and P' is of order $n + 1$, we set $P' = P + e$ in Theorem 2. Then in using the notation of Figure 2, together with the substitution $d = d_1 d_2$ and $f = f_1 f_2$, we use the regular expressions ρ_{uv} of the flowgraph P to construct

$$\begin{aligned} \rho'_{uv} &= \rho_{uv} + \rho_{ux}(d_1 e f_2 * \rho_{yx})\rho_{yv} \\ &\quad -\{\text{paths from } u \text{ to } v \text{ in } P\} \cup \{\text{paths through } e\} \\ &\quad = \{\text{paths from } u \text{ to } v \text{ in } P'\}, \end{aligned}$$

as required.

STRUCTURED PROGRAMS SUFFICE. The notion of a *structured program* (one employing only the sequence, selection, and repetition constructs [2]) is fundamental to what is now, thanks to Dijkstra [3], considered "good" programming practice. But the implementation of primes of order $n > 1$ requires the use of goto statements in those programming languages that otherwise support only the sequence, selection, and repetition. Now according to the Böhm-Jacopini theory, it is always possible to *restructure* an arbitrary given flowgraph so as to avoid the use of such higher-order primes (and to thereby circumvent the need for employing goto statements). That is the conclusion that we come to here.

Theorem 3. *Corresponding to each program flowgraph F is a regular expression ρ_F with*

$$L(F) = \lambda(\rho_F).$$

Proof: By Theorem 1 we have

$$F = P_1 \circ \cdots \circ P_i[e_{j_i} \to F_{ij}] \circ \cdots \circ P_k$$

and with an inductive hypothesis that the theorem is true for flowgraphs of lesser order than that of F, there are regular expressions ρ_{ij} with $L(F_{ij}) = \lambda(\rho_{ij})$. Then according to the Corollary to Theorem 2, there are also regular expressions ρ_{P_i} with $L(P_i) = \lambda(\rho_{P_i})$ on an alphabet of edges-as-symbols. We define

$$\rho_F = \prod_{i=1}^{k} \rho_{P_i}(e_{j_i}(\rho_{ij}))$$

and in combining all of our earlier results, we obtain:

$$\begin{aligned} L(F) &= L(P_1 \circ \cdots \circ P_i[e_{j_i} \to F_{ij}] \circ \cdots \circ P_k) && \text{Theorem 1} \\ &= \prod_{i=1}^{k} L(P_i[e_{j_i} \to F_{ij}]) && \text{Corollary, Theorem 1} \\ &= \prod_{i=1}^{k} L(P_i)(e_{j_i}{:}L(F_{ij})) && \text{Rule (2)} \\ &= \prod_{i=1}^{k} \lambda(\rho_{P_i})(e_{j_i}{:}L(F_{ij})) && \text{Corollary, Theorem 2} \end{aligned}$$

$$= \prod_{i=1}^{k} \lambda(\rho_{P_i})(e_{j_i} : \lambda(\rho_{ij})) \qquad \text{Inductive hypothesis}$$

$$= \prod_{i=1}^{k} \lambda\big(\rho_{P_i}(e_{j_i}(\rho_{ij}))\big) \qquad \text{Lemma}$$

$$= \lambda\left(\prod_{i=1}^{k} \rho_{P_i}(e_{j_i}(\rho_{ij}))\right) \qquad \text{Definition (ii)}$$

$$= \lambda(\rho_F) \qquad \text{Definition of } \rho_F$$

as required.

Corollary (Böhm-Jacopini). *Corresponding to each program flowgraph F is a structured program flowgraph with the same computation set.*

Proof: There is a one-to-one correspondence (see the opening lines of the proof of the Corollary to Theorem 2) between the three binary operations of the regular expression algebra and the three structured programming constructs–sequence, selection (D_1), and repetition (D_2).

Our result is not an "exact replica" of the Böhm-Jacopini theorem: "Given any program, there is an 'equivalent' structured program" since we have used a more elementary semantics than was their intent. In the true Böhm-Jacopini sense, two programs are considered to be *equivalent* only if they compute the same (partial) function, relative to certain well-defined input and output program variables. Since we have "abstracted away" all references to the exact nature of individual program decisions and the specific details of the elementary processes of an algorithm, in fact—we have no variables whatsoever, the idea of a program "computing a function" is not meaningful in our setting. And it is precisely for this reason that we are able to describe the "meaning" of a program (its computation set) at the lowest automata-theoretic linguistic level, that of regular languages. The more extensive semantics would take us to the level of the languages accepted by Turing machines **[8]**, where the whole argument is more complex. Nevertheless, we are confident that our construction (that of Theorem 3 and the Corollary to Theorem 2) represents a "restructuring" process that could easily be transformed in the more extensive semantic domain, in order to obtain the corresponding result in the Böhm-Jacopini sense of program equivalence.

REFERENCES

1. C. Böhm and G. Jacopini, Flow diagrams, Turing machines and languages with only two formation rules, *Communications of the ACM* **9** (1966), 366–371.
2. O. Dahl, E. Dijkstra and C. Hoare, *Structured Programming*, Academic Press, New York, 1972.
3. E. Dijkstra, Goto statement considered harmful, *Communications of the ACM* **11** (1968), 147–148.
4. N. Fenton, *Software Metrics: A Rigorous Approach*, Chapman and Hall, New York, 1991.
5. N. Fenton and R. Whitty, Axiomatic approach to software metrication, *The Computer Journal* **29** (1986), 330–339.
6. F. Harary, *Graph Theory*, Addison-Wesley, Reading, MA, 1972.
7. A. Ginzberg, *Algebraic Theory of Automata*, Academic Press, New York, 1968.
8. B. Marion, Turing machines and computational complexity, *Amer. Math. Monthly* **101** (1994), 61–65.
9. C. McGeoch, Data compression, *Amer. Math. Monthly* **100** (1993), 493–497.
10. R. Prather and S. Guilieri, Decomposition of flowchart schemata, *The Computer Journal* **24** (1981), 258–262.
11. R. Prather, Design and analysis of hierarchical software metrics, *ACM Computing Surveys* **27** (1995), 499–518.

12. A. Salomaa, Two complete axiom systems for the algebra of regular events, *Journal of the ACM* **13** (1966), 158–169.
13. A. Salomaa, *Formal Languages*, Academic Press, New York, 1973.
14. R. Tennent, *Principles of Programming Languages*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

Department of Computer Science
Trinity University
San Antonio, TX 78212
rprather@trinity.edu

News Items from 100 Years Ago in the MONTHLY . . .

The meeting at Zurich, August 9th–11th, of the International Congress of Mathematicians was in every way a success. More than two hundred members took part. America sent seven representatives, including, however, three Cambridge graduates, now transplanted to Pennsylvania, Professors Harkness, Morley, and Charlotte Scott. The greatest mathematician in the world, Sophus Lie, was not expected; and the greatest French mathematician, Poincaré, though down for a speech, did not come; but the actual program was particularly rich and interesting. [p. 229]

The articles on "Euclidean Geometry Without Disputed Axioms," and "Zero, Infinitesimals, Infinity, and the Fundamental Symbol of Indetermination" are published at the request of the authors. They invite criticism on their respective articles, and, if there is any defect in the reasoning by which they arrived at their conclusions, they desire to have the same pointed out. [p. 259]

The degree of Doctor of Philosophy was conferred June 9th, by the University of Pennsylvania, on Prof. Robert. J. Aley, the subject of his thesis being "Some Contributions to the Geometry of the Triangle." We congratulate Dr. Aley on having received this degree as it is not an honorary one, but was attained by actual work done at the University during the past year. [p. 194]

We are in correspondence with several excellent mathematicians who are desirous of securing better positions for next year. If any of our readers know of such positions which are vacant or likely to be come vacant at the end of this school year, we shall be pleased to refer them to these gentlemen. [p. 33]

Dr. L. E. Dickson, who spent last year at the Universities of Göttingen and Paris, has been elected Assistant Professor of Mathematics in the University of California. [p. 230]

We shall be pleased to pay 25 cents each for a limited number of copies of No. 6, Vol. I, and No. 11, Vol. II, of the MONTHLY. Any of our readers wishing to dispose of these numbers should write to us. [p. 63]

. . . Vol. 4, 1897.

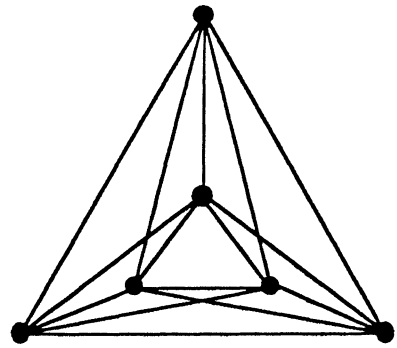


Figure 2

crossings, which turns out to be optimal. Since the complete graphs have a very special structure indeed, we can hope to calculate their crossing numbers.

There are conjectures for the crossing numbers of both the complete and complete bipartite graphs [3]:

$$cr(K_n) = \frac{1}{4}\left\lfloor\frac{n}{2}\right\rfloor\left\lfloor\frac{n-1}{2}\right\rfloor\left\lfloor\frac{n-2}{2}\right\rfloor\left\lfloor\frac{n-3}{2}\right\rfloor$$

and

$$cr(K_{m,n}) = \left\lfloor\frac{m}{2}\right\rfloor\left\lfloor\frac{m-1}{2}\right\rfloor\left\lfloor\frac{n}{2}\right\rfloor\left\lfloor\frac{n-1}{2}\right\rfloor$$

However, these remain open. Some partial results are known: the former has been verified for $n \leq 10$, while the latter holds for $m \leq 6$ and all n [4] and for $m = 7$ and $n \leq 10$ [7].

The best known drawings of $K_{m,n}$ and K_n achieve these values. The description of such a drawing for $K_{m,n}$ is quite simple. Divide both the m-set and the n-set into two as-equal-as-possible parts. Place the m along the y-axis, with half above the x-axis and half below. Similarly, place the n along the x-axis, with half to the left of the y-axis and half to the right. Now join the m to the n using straight lines. The second drawing in Figure 1 is such a drawing of $K_{3,4}$.

Turan's story suggests a variant of the crossing number problem for complete bipartite graphs: find the smallest number of crossings in a *cylindrical* drawing of $K_{n,n}$, that is a drawing of $K_{n,n}$ on a cylinder such that each class of n vertices is on one of the two boundaries of the cylinder.

One way to get a drawing of K_{2n} in the plane is start with a cylindrical drawing of $K_{n,n}$ and then use the top and bottom of the cylinder to complete the drawing of K_{2n}. See Figure 3 for the case $n = 4$.

Obviously, this drawing of K_{2n} has $2\binom{n}{4}$ more crossings than the cylindrical drawing of $K_{n,n}$; this type of drawing of K_{2n} is described in [6]. With an appropriate choice of cylindrical drawing of $K_{n,n}$, the conjectured crossing number of K_{2n} is obtained this way.

One might hope that some better cylindrical drawing of $K_{n,n}$ exists and, therefore, a better drawing of K_{2n} would result. In Section 2, we associate a quadratic form with such drawings. Minimizing the quadratic form, we find the best cylindrical drawing of $K_{n,n}$, and so get the best drawing of K_{2n} *of this type*.

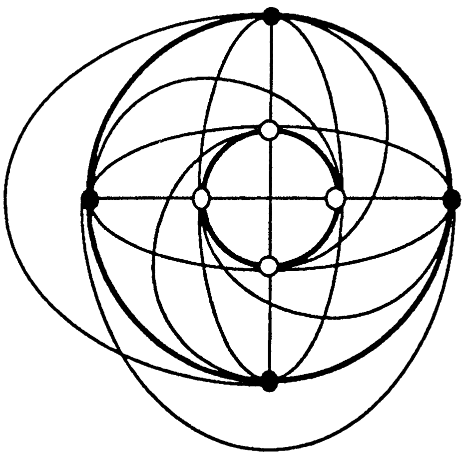


Figure 3

In Section 3 we shall discuss asymptotic values of the crossing numbers of K_n and $K_{n,n}$. It is easy to see (and will be discussed in Section 3) that the sequences $cr(K_n)/\binom{n}{4}$ and $cr(K_{n,n})/\binom{n}{2}^2$ are monotonically increasing and each term is less than 1. Therefore, the limits

$$\lim_{n\to\infty} \frac{cr(K_n)}{\binom{n}{4}} \quad \text{and} \quad \lim_{n\to\infty} \frac{cr(K_{n,n})}{\binom{n}{2}^2}$$

both exist and are at most 1. The conjectures on the values of the crossing numbers $cr(K_n)$ and $cr(K_{n,n})$ imply that the limits are $3/8$ and $1/4$, respectively. We prove in Section 3 that the latter implies the former.

2. CYLINDRICAL DRAWINGS OF $K_{n,n}$. We want to determine a lower bound on the number of crossings in any cylindrical drawing of $K_{n,n}$. We need to discover just what forces a crossing in the drawing. Consider, first, a single vertex v of $K_{n,n}$. All the edges incident with v are drawn across the cylinder to vertices on the other boundary. No two of these edges cross in an optimal drawing; see Figure 4.

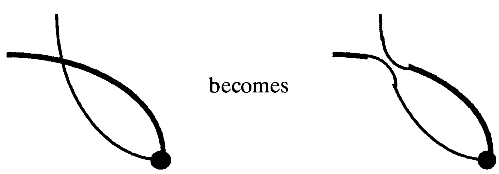


Figure 4

Now consider two vertices v and w on the same boundary. There are several possibilities for how the edges incident with these vertices are drawn, but we can see (Figure 5) that no two edges cross more than once in an optimal drawing. So how can two edges be forced to cross?

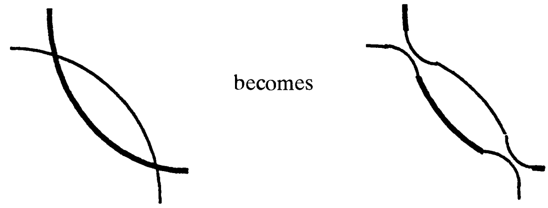


Figure 5

A little thought yields a simple observation.

For each vertex i on the inside boundary, there is a vertex $x_i \in \{1,2,\ldots,n\}$ on the outside boundary such that the simple closed curve consisting of the edges from i to each of x_i and $x_i + 1$ (the arithmetic being taken modulo n), together with the little segment of the outer boundary of the cylinder joining x_i and $x_i + 1$ bounds a disc containing the inner boundary of the cylinder.

As examples, in Figure 6, $x_1 = 5$ and $x_2 = 7$.

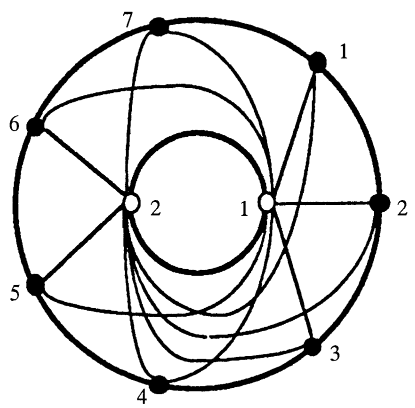


Figure 6

Now it is a simple matter to get a lower bound on the number of crossings given that the values of $x_1, x_2, \ldots, x_n$ are known. We need only deal with these in pairs, i.e., it suffices to calculate the number of crossings among edges incident with the vertices i and j on the inside boundary. If we pick two vertices r and s between $x_i + 1$ and x_j, say, then, among the four edges with ends i or j and r or s, there must be at least one crossing (Figure 7a). Similarly, if r and s are both between $x_j + 1$ and x_i. But if one is between $x_i + 1$ and x_j and the other is between $x_j + 1$ and x_i, then there need not be a crossing (Figure 7b).

Assuming that $1 \le x_i \le x_j \le n$, it follows that there are at least

$$\binom{x_j - x_i}{2} + \binom{n + x_i - x_j}{2}$$

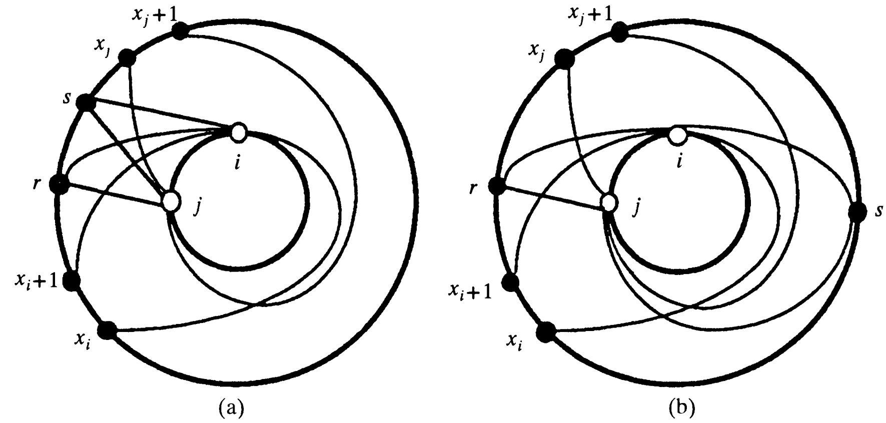


Figure 7

crossings in the drawing among edges incident with i and j. Therefore, a lower bound for the total number of crossings in the drawing is

$$\sum_{1 \le i < j \le n} \binom{|x_j - x_i|}{2} + \binom{n - |x_j - x_i|}{2}.$$

Using the relation $\binom{y}{2} = y(y-1)/2$, we see that the lower bound is the function

$$f(x_1, x_2, \ldots, x_n) = \binom{n}{2}^2 + \Big(\sum_{1 \le i < j \le n} |x_i - x_j|^2 \Big) - n \Big(\sum_{1 \le i < j \le n} |x_i - x_j| \Big).$$

Ordering the variables so $1 \le x_1 \le x_2 \le \cdots \le x_n \le n$, we see that the lower bound is given by the quadratic function

$$F(x_1, x_2, \ldots, x_n) = \binom{n}{2}^2 + \Big(\sum_{1 \le i < j \le n} (x_j - x_i)^2 \Big) - n \Big(\sum_{1 \le i < j \le n} (x_j - x_i) \Big).$$

Clearly F has a minimum, which we shall determine.

The function F is differentiable and

$$\frac{\partial F}{\partial x_i} = 2 \sum_{j \ne i} (x_i - x_j) + n(n - 2i + 1) = 2nx_i - 2 \sum_{j=1}^{n} x_j + n(n - 2i + 1).$$

Setting $S = \sum_{j=1}^{n} x_j$ and $\nabla F = 0$, we find that

$$x_i = \frac{2S - n(n - 2i + 1)}{2n}$$

It is an easy calculation to see that $x_{i+1} - x_i = 1$ and, therefore, setting $x_i = i$ yields a solution to these equations. Moreover, every other solution is obtained from this one by adding the same quantity t to each x_i.

This means that there is an integral minimum for F, namely $x_i = i$, $i = 1, 2, \ldots, n$. Thus, a lower bound for the number of crossings in a cylindrical

drawing of $K_{n,n}$ is

$$\begin{aligned}F(1,2,\ldots,n) &= \sum_{1\le i<j\le n}\binom{j-i}{2} + \sum_{1\le i<j\le n}\binom{n-j+i}{2}\\ &= \sum_{k=1}^{n}\binom{k}{2}(n-k) + \sum_{k=1}^{n-1}\binom{k}{2}k\\ &= n\sum_{k=1}^{n-1}\binom{k}{2} = n\binom{n}{3}.\end{aligned}$$

This is attainable; see Figure 3 for the case $n = 4$. The drawing of K_{2n} obtained from this optimal cylindrical drawing of $K_{n,n}$ has the same number of crossings as the conjectured crossing number of K_{2n}.

3. ASYMPTOTICS. The following classical counting argument estimates the crossing number of K_{n+1} in terms of the crossing number of K_n. Deleting in turn each vertex from a drawing of K_{n+1} yields $n + 1$ different drawings of K_n. Each of these must have at least $cr(K_n)$ crossings, so we estimate the crossing number of K_{n+1} by $(n + 1)cr(K_n)$.

How many times do we count a given crossing? A given crossing from K_{n+1} occurs in one of the drawings of K_n if the four vertices that are the ends of the edges involved in the crossing are all in the K_n we pick. Given that we must have these four vertices, there are $n - 4$ vertices left to be picked from the remaining $n - 3$ vertices of the K_{n+1}. Thus, the four vertices (and so the particular crossing) are in $n - 3$ of the K_n. Thus, each crossing is counted $n - 3$ times and we have the estimate

$$cr(K_{n+1}) \ge \frac{n+1}{n-3}\,cr(K_n).$$

This estimate is equivalent to

$$\frac{cr(K_{n+1})}{\binom{n+1}{4}} \ge \frac{cr(K_n)}{\binom{n}{4}}.$$

Therefore, the sequence $cr(K_n)/\binom{n}{4}$ is nondecreasing. Since it is bounded above by 1, it has a limit, say LC (for Limit of Complete graphs).

An entirely analogous argument shows that $cr(K_{n,n})/\binom{n}{2}^2$ has a limit LB (for Limit of complete Bipartite graphs). The drawings of $K_{n,n}$ such as the second drawing in Figure 1 show that $LB \le 1/4$.

It is easy to see that the conjectures as to the crossing numbers for K_n and $K_{m,n}$ imply that $LC = 3/8$ and $LB = 1/4$. We now show there is a relation between these limits.

Theorem. $LC \ge (3/2)LB$. *If* $LB = 1/4$, *then* $LC = 3/8$.

Proof: Let K_{2n} be drawn with $cr(K_{2n})$ crossings. Within this drawing, there are many different drawings of $K_{n,n}$. We need to estimate how many drawings of $K_{n,n}$ there are and how many of these contain a given crossing.

We shall count *ordered* $K_{n,n}$'s, i.e., those where we first pick one set of n and then the other set of n. There are, evidently, $\binom{2n}{n}$ such graphs.

Now consider a given crossing involving the edges ab and cd of K_{2n}. One of a and b must be in the first set of n chosen, and similarly for c and d. Thus, there are 4 ways to distribute a, b, c, d into the first set of n chosen, if this crossing is to occur in the resulting $K_{n,n}$. There are $2n - 4$ vertices left, of which $n - 2$ are to be put into the first set of n. Therefore, there are $4\binom{2n-4}{n-2}$ different $K_{n,n}$'s that contain the given crossing, and hence

$$cr(K_{2n}) \geq \frac{\binom{2n}{n}}{4\binom{2n-4}{n-2}} cr(K_{n,n}).$$

Divide both sides of this inequality by $\binom{2n}{4}$ and do some easy arithmetic to get

$$\frac{cr(K_{2n})}{\binom{2n}{4}} \geq \frac{3}{2} \frac{cr(K_{n,n})}{\binom{n}{2}^2}.$$

Now taking the limit as n tends to infinity, we have the relation

$$LC \geq (3/2) LB.$$

It follows that if $LB = 1/4$, then $LC \geq 3/8$. Since we have previously noted $LC \leq 3/8$, it follows that if $LB = 1/4$, then $LC = 3/8$. ■

This theorem shows that the conjecture for $cr(K_{n,n})$ implies the conjecture for $cr(K_{2n})$, at least asymptotically. Does the converse hold?

Probably this cannot be derived by counting. The reason why the proof of the theorem works (as the proof shows!) is that *any* (almost) optimal drawing of K_{2n} contains a drawing of $K_{n,n}$ that is economical in the sense that it has (almost) as few crossings as the conjectured value for $cr(K_{n,n})$.

For the converse, however, we do not know of a natural way to extend (almost) optimal drawings of $K_{n,n}$ to economical drawings of K_{2n}. The optimal cylindrical drawings of $K_{n,n}$ have many more than $cr(K_{n,n})$ crossings.

REFERENCES

1. M. R. Garey and D. S. Johnson, Crossing number is NP-complete, *SIAM J. Alg. Disc. Meth.* **4** (1983), 312–316.
2. R. K. Guy, The decline and fall of Zarankiewicz's theorem in *Proof Techniques in Graph Theory*, Academic Press, New York, 1969, pp. 63–69.
3. F. Harary, *Graph Theory*, Addison Wesley, Reading, MA, 1972.
4. D. J. Kleitman, The crossing number of $K_{5,n}$, *J. Combin. Theory* **9** (1970), 315–322.
5. P. Turán, A note of welcome, *J. Graph Theory* **1** (1977), 7–9.
6. A. T. White, Topological graph theory, in *Selected Topics in Graph Theory*, Academic Press, New York.
7. D. R. Woodall, Cyclic-order graphs and Zarankiewicz's crossing-number conjecture, *J. Graph Theory* **17** (1993), 657–671.

Department of Mathematics and Statistics
Carleton University
Ottawa Canada K1S 5B6
brichter@math.carleton.ca

Mathematics Institute
The Technical University of Denmark
2800 Lynngby Denmark

Carries, Combinatorics, and an Amazing Matrix

John M. Holte

This is the story of a serendipitous discovery. It began when I was investigating a mundane subject: carries in addition. To my surprise, a probabilistic perspective and some heavy-duty number crunching revealed a mathematical cache: an infinite collection of stochastic matrices in every dimension exhibiting an unusual symmetry and multifaceted combinatorial features. For each matrix $\mathbf{\Pi}$:

- The eigenvalues are all positive and form a finite, decreasing, geometric sequence; furthermore, if we diagonalize $\mathbf{\Pi}$ as $\mathbf{U}^{-1}\mathbf{\Pi}\mathbf{U} = \mathbf{D}$, where the eigenvalues are arranged in decreasing order in the diagonal matrix $\mathbf{D}$, then, aside from a constant of proportionality:
- The entries in the row of U^{-1} corresponding to the eigenvalue 1 are Eulerian numbers.
- The entries in the row of $\mathbf{U}^{-1}$ corresponding to the least eigenvalue are the entries in a row of Pascal's triangle, but with alternating signs; the entries in the column of $\mathbf{U}$ corresponding to this eigenvalue are their reciprocals.
- The entries in the first and last rows of $\mathbf{U}$ are respectively unsigned and signed Stirling numbers of the first kind.

These unanticipated relationships first came to light when I explored the territory numerically, using Mathematica®. That started me on a project that cycled through phases of computer experimentation, conjecture, and rigorous mathematics. The mathematics involved included generating functions, recurrence relations, summation and matrix manipulation, combinatorial identities, and discrete probability—the techniques of "concrete mathematics" ([**8**]; see also [**18**]). This article is an invitation to aficionados of concrete mathematics to enjoy a guided tour of some wonderful sights. Along the way we will also point out several interesting side trips (exercises) for explorers.

THE PROBLEM. When we add two long random base-ten (say) numbers, how often do we have a carry (of 1) from one column to the next? For example, consider the following addition of two fifty-digit numbers composed of digits taken from a table of random numbers:

010011	00110	11100	01111	00001	00000	01101	11111	00000	1100
24003	80475	19793	71578	52010	72216	15692	96689	80452	46312
+16129	49245	21693	20946	60874	82351	32516	23823	30046	06870
40133	29720	41486	92525	12885	54567	48209	20513	10498	53182

We observe that we got a carry-out of 0 in 27 cases and a carry-out of 1 in 23 cases, or 54% and 46%, respectively. It would be natural to conjecture that in the long run, as the number of digits increases without bound, the relative frequencies would be 50%–50%. This is true, and a thorough treatment is given in [**12**, pp. 262–263].

What happens if we add *three* long random numbers? When I asked some faculty colleagues, they conjectured that carries of 0, 1, and 2 would be equally likely. In a seminar for students, one participant confidently asserted that there would be mostly 1's. In the following sum of three 50-digit random numbers

111011	10111	11000	10111	10210	11102	11122	01011	11210	2112
05453	03060	83621	43443	07082	04401	15299	64642	73497	38426
67711	70528	46700	00171	55077	11440	95932	91116	17255	19649
76306	39287	31026	49339	70267	68885	98147	70311	43856	37376
149471	12876	61347	92954	32426	84728	09380	26070	34608	95451

we have 12 (24%) carries of 0, 31 (62%) carries of 1, and 7 (14%) carries of 2; perhaps the student was right.

But are these empirical percentages good estimates of the long-run frequencies? And more generally, what is the long-run frequency of each possible carry value when we add any number of long numbers represented in any base?

THE CARRIES PROCESS. Consider the addition of m random n-digit base-b numbers:

Carries	C_n	C_{n-1}	C_{n-2}	$\cdots$	C_2	C_1	$C_0 = 0$
Addends		$X_{1,n-1}$	$X_{1,n-2}$	$\cdots$	$X_{1,2}$	$X_{1,1}$	$X_{1,0}$
		$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
	$+$	$X_{m,n-1}$	$X_{m,n-2}$	$\cdots$	$X_{m,2}$	$X_{m,1}$	$X_{m,0}$
Sum	S_n	S_{n-1}	S_{n-2}	$\cdots$	S_2	S_1	S_0

We assume that the $\{X_{h,k}\}$ are independent uniformly distributed random digits. The key to our analysis is this probabilistic insight: *the carries form a finite Markov chain*:

$$\Pr(C_{k+1} = c_{k+1} | C_k = c_k, \dots, C_1 = c_1, C_0 = 0) = \Pr(C_{k+1} = c_{k+1} | C_k = c_k).$$

This is true because the carry-out C_{k+1} depends only on C_k and, of course, the digits $X_{1,k}$, $X_{2,k}, \dots,$ and $X_{m,k}$.

What are the possible values of C_k? Those who have experience with adding long columns of figures by hand know that the carry-out can be anything from 0 to $m-1$.[1] Thus the state space of the carries process $\langle C_k \rangle$ is $\{0, 1, \dots, m-1\}$. Furthermore, it is possible to get from any state to any other state in $\lfloor \log_b(m-1) \rfloor + 1$ steps. A probabilist would say that this Markov chain is acyclic (aperiodic) and irreducible.

Let $\mathbf{\Pi} = [\pi_{ij}]$ denote the transition matrix:

$$\pi_{ij} = \Pr(\text{carry-out} = j | \text{carry-in} = i) \quad \text{where } 0 \le i, j \le m-1.$$

Because the states of the Markov chain are numbered $0, \dots, m-1$, we number the rows and columns of $\mathbf{\Pi}$ in the same way. Now, to calculate π_{ij}, consider the base-b addition in the kth place:

$$C_{k+1} = j \Leftrightarrow jb \le i + X_{1,k} + \cdots + X_{m,k} < (j+1)b$$

where $0 \le X_{1,k}, \dots, X_{m,k} \le b-1$. Introducing the slack variable Y, we observe

[1]An interesting induction problem is to prove that the maximum possible value of the carry C_k is $m - 1 - \lfloor (m-1)/b^k \rfloor$.

that this is equivalent to

$$X_{1,k} + \cdots + X_{m,k} + Y = (j+1)b - 1 - i =: z \tag{1}$$

where $0 \le X_{1,k}, \ldots, X_{m,k}, Y \le b - 1$. As Tucker [**16**, p. 311] notes concerning a similar problem, "By using generating functions to solve this problem, we [do] not need to know anything about the inclusion-exclusion complexities of this problem. Generating functions automatically [perform] the required combinatorial logic!" So now we invoke generating functions (and we gear up to the level of chapter 6 of [**16**] or chapter 2 of *generatingfunctionology* [**18**]). The number of integer solutions of (1) is the same as the coefficient of x^z in $(1 + x + x^2 + \cdots + x^{b-1})^{m+1}$. Because

$$(1 + x + x^2 + \cdots + x^{b-1})^{m+1} = (1 - x^b)^{m+1}(1 - x)^{-(m+1)}$$

and

$$(1 - x^b)^{m+1} = \sum_r \binom{m+1}{r}(-x^b)^r$$

and

$$(1 - x)^{-(m+1)} = \sum_{s \ge 0} \binom{m+s}{m} x^s,$$

the desired coefficient is

$$\sum_{r \le z/b} (-1)^r \binom{m+1}{r}\binom{m+z-rb}{m}.$$

Since $r \le z/b = j + 1 - (i+1)/b$ if and only if $r \le j - \lfloor i/b \rfloor$, we may summarize our result as follows.

Theorem 1. *The carries process $\langle C_k \rangle$ for the base-b addition of m random numbers is a finite Markov chain with state space $\{0, 1, \ldots, m-1\}$ and transition matrix $\boldsymbol{\Pi} = [\pi_{ij}]$ given by*

$$\pi_{ij} = b^{-m} \sum_{r=0}^{j - \lfloor i/b \rfloor} (-1)^r \binom{m+1}{r}\binom{m - 1 - i + (j+1-r)b}{m}.$$

When $b = 2$, the number of bit-valued solutions of (1) is simply $\binom{m+1}{z}$, so

$$\pi_{ij} = 2^{-m}\binom{m+1}{2j - i + 1} \quad \text{in the binary case.}$$

Let's look at some other examples. When $b = 10$ and $m = 2, 3, 4$, then $\boldsymbol{\Pi}$ is

$$\begin{bmatrix} 0.55 & 0.45 \\ 0.45 & 0.55 \end{bmatrix}, \quad \begin{bmatrix} 0.220 & 0.660 & 0.120 \\ 0.165 & 0.670 & 0.165 \\ 0.120 & 0.660 & 0.220 \end{bmatrix}, \quad \begin{bmatrix} 0.0715 & 0.5280 & 0.3795 & 0.0210 \\ 0.0495 & 0.4840 & 0.4335 & 0.0330 \\ 0.0330 & 0.4335 & 0.4840 & 0.0495 \\ 0.0210 & 0.3795 & 0.5280 & 0.0715 \end{bmatrix}.$$

The 0.0210 in the upper right corner, for example, signifies that, given a carry-in of 0 to a column of 4 random decimal digits, the probability of a carry-out of 3 is 0.0210. For a general base b we obtain the following formulas when $m = 2, 3$:

$$\boldsymbol{\Pi} = \frac{1}{2b}\begin{bmatrix} b+1 & b-1 \\ b-1 & b+1 \end{bmatrix} \quad \text{and}$$

$$\boldsymbol{\Pi} = \frac{1}{6b^2}\begin{bmatrix} b^2 + 3b + 2 & 4b^2 - 4 & b^2 - 3b + 2 \\ b^2 - 1 & 4b^2 + 2 & b^2 - 1 \\ b^2 - 3b + 2 & 4b^2 - 4 & b^2 + 3b + 2 \end{bmatrix}.$$

CROSS-SYMMETRY. These examples reveal that Π has an unusual sort of symmetry: it is radially symmetric about its center. A typical crossword puzzle grid has the same sort of symmetry. This symmetry is familiar to matrix theorists, who call it "centrosymmetry," and to statisticians, who call it "cross-symmetry." See **[17]** for a survey.

Theorem 2. *For $i, j = 0, 1, \ldots, m-1$, we have $\pi_{m-1-i, m-1-j} = \pi_{i,j}$, i.e.,*

$$\Pr(C_{k+1} = m-1-j | C_k = m-1-i) = \Pr(C_{k+1} = j | C_k = i).$$

Proof: The cross-symmetry is not obvious from the formula in Theorem 1, so we turn to the probabilistic definition. Given that $C_k = i$, we have $C_{k+1} = j$ if and only if

$$jb \le i + X_{1,k} + X_{2,k} + \cdots + X_{m,k} \le (j+1)b - 1. \tag{2}$$

The $\{X_{h,k}\}$ are independent random variables that are uniformly distributed on $\{0, 1, \ldots, b-1\}$. Accordingly, the equation $\tilde{X}_{h,k} := b - 1 - X_{h,k}$ defines independent random variables that are also uniformly distributed on $\{0, 1, \ldots, b-1\}$. Now if we negate the inequalities (2) and add $mb - 1$, we get

$$(m-1-j+1)b - 1 \ge m - 1 - i + \tilde{X}_{1,k} + \tilde{X}_{2,k} + \cdots + \tilde{X}_{m,k} \ge (m-1-j)b,$$

which is the condition for $C_{k+1} = m-1-j$ given that $C_k = m-1-i$. ∎

EIGENVALUES AND EIGENVECTORS AND SERENDIPITY. Let's return to the carries problem. It is well known in Markov chain theory that our original question concerning the long-run relative frequencies of the carry values is answered by the stationary probability vector, i.e., the row vector $\mathbf{v} = (p_0, \ldots, p_{m-1})$ with nonnegative entries summing to 1 that satisfies $\mathbf{v}\Pi = \mathbf{v}$. Thus, $\mathbf{v}$ is the left eigenvector of Π associated with the eigenvalue 1. When I used Mathematica® to calculate some sample cases, out of curiosity I asked for more than $\mathbf{v}$ alone; I asked for the entire eigensystem. That's when I discovered the surprises hidden in the matrix Π.

Let's look at the eigenvalues first. For $b = 10$ and $m = 2, 3, 4, 5$, we find these eigenvalue sets: {1, 0.1}, {1, 0.1, 0.01}, {1, 0.1, 0.01, 0.001}, and {1, 0.1, 0.01, 0.001, 0.0001}. For $b = 2$ and $m = 5$ we get $\{1, 1/2, 1/4, 1/8, 1/16\}$.

Conjecture 1. The eigenvalues of Π are given by the geometric sequence $1, b^{-1}, \ldots, b^{-(m-1)}$.

The eigenvectors for the two $m = 5$ cases ($b = 10$ and $b = 2$) turn out to be the same. Further numerical experimentation shows that the eigenvectors are independent of the base!

Conjecture 2. The eigenvectors do *not* depend on b.

What do these eigenvectors look like? If we assemble these (row) eigenvectors in a matrix $\mathbf{V} = [v_{ij}] = [v_{ij}(m)]$ for m = 2, 3, 4, and 5, we get:

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 4 & 1 \\ 1 & 0 & -1 \\ 1 & -2 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 11 & 11 & 1 \\ 1 & 3 & -3 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -3 & 3 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 26 & 66 & 26 & 1 \\ 1 & 10 & 0 & -10 & -1 \\ 1 & 2 & -6 & 2 & 1 \\ 1 & -2 & 0 & 2 & -1 \\ 1 & -4 & 6 & -4 & 1 \end{bmatrix}. \tag{3}$$

Familiar sequences emerge at the bottom and top of **V**.

Conjecture 3. The bottom row of **V** is proportional to a row of Pascal's triangle, but with alternating signs.
Conjecture 4. The top row of **V** is proportional to a row of Eulerian numbers.

EULERIAN NUMBERS. The first few Eulerian numbers are listed in the following table.

n	$\left\langle {n \atop 0} \right\rangle$	$\left\langle {n \atop 1} \right\rangle$	$\left\langle {n \atop 2} \right\rangle$	$\left\langle {n \atop 3} \right\rangle$	$\left\langle {n \atop 4} \right\rangle$	$\left\langle {n \atop 5} \right\rangle$	$\left\langle {n \atop 6} \right\rangle$
0	1						
1	1	0					
2	1	1	0				
3	1	4	1	0			
4	1	11	11	1	0		
5	1	26	66	26	1	0	

It appears that $v_{0j}(m) = \left\langle {m \atop j} \right\rangle$ for $j = 0, \dots, m-1$. The Eulerian numbers, first discussed by Euler (of course) in [**6**, pp. 485–487], [**7**, pp. 373–375], arise naturally in the study of random permutations; see [**13**, sect. 5.1.3] and the references there, [**2**], [**4**], [**5**, ch. 10], [**14**], and [**15**]. They satisfy the recurrence relation (see [**8**, sect. 6.2])

$$\left\langle {n \atop k} \right\rangle = (k+1)\left\langle {n-1 \atop k} \right\rangle + (n-k)\left\langle {n-1 \atop k-1} \right\rangle \quad \text{for integer } n > 0 \tag{4}$$

with the boundary condition $\left\langle {0 \atop k} \right\rangle = \delta_{0k}$, the Kronecker delta. Using this relation and induction, one may deduce (as in [**3**])

$$\sum_k \left\langle {n \atop k} \right\rangle = n!. \tag{5}$$

Anticipating the verification of Conjecture 4, we normalize the Eulerian numbers in accordance with (5) to get *the stationary probabilities for the carries process*:

$$p_j = \frac{1}{m!}\left\langle {m \atop j} \right\rangle \quad \text{for } j = 0, \dots m-1.$$

In particular, the long-run relative frequencies of carry values are $(\frac{1}{2}, \frac{1}{2})$ for $m = 2$ and $(\frac{1}{6}, \frac{2}{3}, \frac{1}{6})$ for $m = 3$. We see that our empirical values came reasonably close.

The explicit formula

$$\left\langle {n \atop k} \right\rangle = \sum_{r=0}^{k} (-1)^r \binom{n+1}{r} (k+1-r)^n \tag{6}$$

was given by Euler himself. Notation for Eulerian numbers is not standardized; our notation conforms to that of [**8**].

THE EULERIAN RECURRENCE AND V. How can we find an explicit formula for **V**, the matrix whose rows are the left eigenvectors of **Π**? If we are clever or lucky, we can guess the right answer and then verify it.

Playing with the **V** cases in (3), we find that the Eulerian recurrence (4) holds also for *every* row of **V**, i.e.,

$$v_{ij}(m) = (j+1)v_{ij}(m-1) + (m-j)v_{i,j-1}(m-1) \quad \text{for } 0 \le i < m, \tag{7}$$

where we define $v_{i,-1}(m) = 0$ and $v_{im}(m) = 0$. This recurrence cannot give us the

last row of $\mathbf{V}(m)$ in terms of $\mathbf{V}(m-1)$, because the latter matrix is short one row. But if Conjecture 3 is correct, we can paste in the last row of $\mathbf{V}(m)$ by the formula

$$v_{m-1,j}(m) = (-1)^j\binom{m-1}{j}.$$

A little calculation shows that $\left|{m \atop j}\right| := (-1)^j\binom{m-1}{j}$ has the near-Eulerian property

$$\left|{m \atop j}\right| + (j+1)\left|{m \atop j}\right| + (m-j)\left|{m \atop j-1}\right|,$$

so (7) will be satisfied for $i = m-1$ if we define $v_{m-1,j}(m-1) = (-1)^j\binom{m-1}{j}$, or,

$$v_{mj}(m) = (-1)^j\binom{m}{j}. \tag{8}$$

This is an equation for the row *below* the bottom row of $\mathbf{V}$. Now, to see the pattern that generalizes, we make the nonobvious observation [1, p. 822] that

$$(-1)^j\binom{m}{j} = \binom{m+1}{0} - \binom{m+1}{1} + \cdots + (-1)^j\binom{m+1}{j}.$$

This equation, (8), Conjecture 4, and (6) give

$$v_{mj}(m) = \sum_{r=0}^{j}(-1)^r\binom{m+1}{r}(j+1-r)^0, \text{ and}$$

$$v_{0j}(m) = \sum_{r=0}^{j}(-1)^r\binom{m+1}{r}(j+1-r)^m,$$

so we conjecture that

$$v_{ij} = v_{ij}(m) = \sum_{r=0}^{j}(-1)^r\binom{m+1}{r}(j+1-r)^{m-i}. \tag{9}$$

Theorem 3. *Let* $\mathbf{V} = [v_{ij}]$ *be the* $m \times m$ *matrix given by* (9) *for* $0 \le i, j < m$, *and let* $\mathbf{D} = \operatorname{diag}\{1, b^{-1}, \ldots, b^{-m-1}\}$. *Then*

$$\mathbf{V\Pi V}^{-1} = \mathbf{D}.$$

Assuming Theorem 3, we have $\mathbf{\Pi} = \mathbf{V}^{-1}\mathbf{DV}$; this is an equation that may be used to define $\mathbf{\Pi} = \mathbf{\Pi}(b)$ for every complex $b \ne 0$ and to prove $\mathbf{\Pi}(ab) = \mathbf{\Pi}(a)\mathbf{\Pi}(b)$ for all nonzero complex numbers a and b. When a and b are bases—say $a = 5$ and $b = 2$, whence $ab = 10$—this may be explained as follows. We may rewrite each base-ten digit T in the mixed-radix system having bases 5 and 2: $T = 0 \times 5 \times 2 + F \times 2 + B$; now when the carry-in i is applied to the binary column, it leads to an intermediate carry of k to the base-5 column with probability $\pi_{ik}(2)$, which then generates a carry-out of j with probability $\pi_{kj}(5)$, and so $\pi_{ij}(10) = \Sigma\pi_{ik}(2)\pi_{kj}(5)$, i.e., $\mathbf{\Pi}(10) = \mathbf{\Pi}(2)\mathbf{\Pi}(5)$.

Proof: Concrete Mathematics Ahead. Here we'll make heavy use of "concrete mathematics" techniques. First we observe that

$$v_{ij} = \sum_{r=0}^{j+1}(-1)^r\binom{m+1}{r}(j+1-r)^{m-i}$$

is the convolution, or Cauchy product, of the sequences (in k) $\left\langle (-1)^k \binom{m+1}{k} \right\rangle$ and $\langle k^{m-i} \rangle$ evaluated at $j+1$, i.e.,

$$v_{ij} = \text{coefficient of } x^{j+1} \text{ in } \sum_{k \geq 0} (-1)^k \binom{m+1}{k} x^k \cdot \sum_{k \geqslant 0} k^{m-i} x^k.$$

Now we use the binomial theorem and the generating function of $\langle k^n \rangle$,

$$\sum_{k \geq 0} k^n x^k = \left(x \frac{d}{dx} \right)^n (1-x)^{-1}, \tag{10}$$

to get

$$v_{ij} = \text{coefficient of } x^{j+1} \text{ in } (1-x)^{m+1} \left(x \frac{d}{dx} \right)^{m-i} (1-x)^{-1}$$

$$= \text{coefficient of } x^j \text{ in } x^{-1} (1-x)^{m+1} \left(x \frac{d}{dx} \right)^{m-i} (1-x)^{-1}.$$

Thus we obtain the generating function of the ith row of $\mathbf{V}$:

$$\sum_{j \geq 0} v_{ij} x^j = x^{-1} (1-x)^{m+1} \left(x \frac{d}{dx} \right)^{m-i} (1-x)^{-1}. \tag{11}$$

When $i = 0$, this generating function is $x^{-1} f_m(x)$, where $f_m(x)$ is the Eulerian polynomial of degree m (see **[14]**, **[4]**).

We must show

$$\sum_{k=0}^{m-1} v_{ik} \pi_{kj} = b^{-i} v_{ij} \text{ for } i, j = 0, 1, \ldots, m-1.$$

By substituting, interchanging the order of summation (an entertaining exercise!), and simplifying, we get

$$\sum_{k=0}^{m-1} v_{ik} \pi_{kj} = b^{-m} \sum_{k=0}^{m-1} \sum_{r=0}^{j - \lfloor k/b \rfloor} (-1)^r \binom{m+1}{r} \binom{m-1-k+(j+1-r)b}{m} v_{ik}$$

$$= b^{-m} \sum_{r=0}^{j} \sum_{k=0}^{(m-1) \wedge ((j+1-r)b-1)} (-1)^r \binom{m+1}{r}$$

$$\times \binom{m-1-k+(j+1-r)b}{m} v_{ik}$$

$$= b^{-m} \sum_{r=0}^{j} (-1)^r \binom{m+1}{r} \sum_{k=0}^{(j+1-r)b-1}$$

$$\times \binom{m-1-k+(j+1-r)b}{m} v_{ik}.$$

Let $K = (j+1-r)b - 1$. The inner sum, $\sum_{k=0}^{K} \binom{m+K-k}{m} v_{ik}$, is the convolution of the sequences (in k) $\left\langle \binom{m+k}{m} \right\rangle$ and $\langle v_{ik} \rangle$ evaluated at K. We know that

$$\sum_{k \geq 0} \binom{m+k}{m} x^k = (1-x)^{-m-1},$$

and we have the generating function of $\langle v_{ik} \rangle$ in (11). Thus, the inner sum is equal to the coefficient of x^K in

$$(1-x)^{-m-1} \cdot x^{-1}(1-x)^{m+1}\left(x\frac{d}{dx}\right)^{m-i}(1+x)^{-1} = x^{-1}\left(x\frac{d}{dx}\right)^{m-i}(1-x)^{-1},$$

which, invoking (10), is

$$\sum_{k=0}^{K}\binom{m+K-k}{m}v_{ik} = (K+1)^{m-i} = ((j+1-r)b)^{m-i}$$

Therefore,

$$\begin{aligned} &b^{-m}\sum_{r=0}^{j}(-1)^r\binom{m+1}{r}\sum_{k=0}^{K}\binom{m+K-k}{m}v_{ik} \\ &= b^{-m}\sum_{r=0}^{j}(-1)^r\binom{m+1}{r}((j+1-r)b)^{m-i} \\ &= b^{-i}\sum_{r=0}^{j}(-1)^r\binom{m+1}{r}(j+1-r)^{m-i} = b^{-i}v_{ij}. \end{aligned}$$

■

CONFIRMATION OF THE CONJECTURES AND MORE. Theorem 3 tells us that the rows of **V** are left eigenvectors of **Π** corresponding to the eigenvalues $1, b^{-1}, \dots, b^{-(m-1)}$, so Conjectures 1 and 2 are true. Also the formula for v_{0j} is the same as the explicit formula (6) for the Eulerian numbers, so Conjecture 4 is true. Finally, letting $i = m-1$ in (11), we get

$$\sum_{j\geq 0} v_{m-1,j}x^j = x^{-1}(1-x)^{m+1}x\frac{d}{dx}(1-x)^{-1} = (1-x)^{m-1},$$

so Conjecture 3 follows, by the binomial theorem.

There are other patterns in **V**. It is easy to verify that the leftmost column of **V** is all 1's. It is a little harder to show that the rightmost column has alternating $+1$'s and -1's, but it is a splendid opportunity to use the calculus of finite differences. Both exercises are left to the reader.

THE RIGHT EIGENVECTOR MATRIX U: EMPIRICAL RESULTS. Let's look at the right eigenvectors of the transition matrix **Π**. As an alternative to direct computation of the eigenvectors, we may compute the inverse of the matrix **V**. Numerical experimentation reveals that, in order to get integer values, we should multiply by $m!$, so we let $\mathbf{U} = m!\mathbf{V}^{-1}$. For $m = 2, 3, 4, 5$, we find that **U** is:

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 3 & 2 \\ 1 & 0 & -1 \\ 1 & -3 & 2 \end{bmatrix}, \quad \begin{bmatrix} 1 & 6 & 11 & 6 \\ 1 & 2 & -1 & -2 \\ 1 & -2 & -1 & 2 \\ 1 & -6 & 11 & -6 \end{bmatrix}, \quad \begin{bmatrix} 1 & 10 & 35 & 50 & 24 \\ 1 & 5 & 5 & -5 & -6 \\ 1 & 0 & -5 & 0 & 4 \\ 1 & -5 & 5 & 5 & -6 \\ 1 & -10 & 35 & -50 & 24 \end{bmatrix}.$$

Tantalizing patterns are already visible in these first few examples. Even though the columns are the eigenvectors, the top and bottom *rows* leap out at the combinatorial *cognoscenti*: They are Stirling numbers of the first kind! The pattern of the eigenvector in the last column may be exposed by dividing by $(m-1)!$: reciprocals of binomial coefficients with alternating signs! Forming difference tables of the columns reveals more patterns: It appears that the jth difference of

the jth column is a constant—$(-1)^n m!/(m-j)!$—which would make the jth column a polynomial of degree j in the row index i. To summarize, for the matrix $\mathbf{U} = m!\mathbf{V}^{-1}$, we propose:

Conjecture 5. Column j is a degree-j polynomial function of row index i.
Conjecture 6. The entries in the final column are proportional to reciprocals of entries in a row of Pascal's triangle with alternating signs.
Conjecture 7. The top row consists of unsigned Stirling numbers of the first kind (in reverse order).
Conjecture 8. The bottom row consists of signed Stirling numbers of the first kind (in reverse order).

STIRLING NUMBERS OF THE FIRST KIND. The first few (unsigned) Stirling numbers of the first kind are as follows.

n	$\left[{n \atop 0}\right]$	$\left[{n \atop 1}\right]$	$\left[{n \atop 2}\right]$	$\left[{n \atop 3}\right]$	$\left[{n \atop 4}\right]$	$\left[{n \atop 5}\right]$	$\left[{n \atop 6}\right]$
0	1						
1	0	1					
2	0	1	1				
3	0	2	3	1			
4	0	6	11	6	1		
5	0	24	50	35	10	1	

The Stirling number $\left[{n \atop k}\right]$ may be characterized combinatorially as the number of ways n objects can be arranged into k cycles, but for our purposes we characterize Stirling numbers algebraically. Rising factorial powers may be represented in terms of ordinary powers by means of unsigned Stirling numbers of the first kind:

$$x^{\overline{n}} = x(x+1)(x+2)\cdots(x+n-1) = \sum_k \left[{n \atop k}\right] x^k; \tag{12}$$

falling factorial powers may be represented in terms of ordinary powers by means of signed Stirling numbers the first kind:

$$x^{\underline{n}} = x(x-1)(x-2)\cdots(x-n+1) = \sum_k (-1)^{n-k} \left[{n \atop k}\right] x^k. \tag{13}$$

(See [**8**, sect. 6.1], [**11**, pp. 65–68], or [**10**, ch. 4].)

THE RIGHT EIGENVECTORS. How can we find an explicit formula for the matrix $\mathbf{U}$ of right eigenvectors of $\mathbf{\Pi}$? One way would be to solve $\mathbf{\Pi U} = \mathbf{UD}$, which appears to be very difficult. My way was to find $\mathbf{U}$ by solving $\mathbf{UV} = m!\mathbf{I}$. It turns out to take longer to solve this equation than it does to prove the answer is right, so let's start with the answer.

Theorem 4. *Let* $\mathbf{V} = [v_{ij}]$ *be the* $m \times m$ *matrix given by* (9). *Then* $m!\mathbf{V}^{-1} = [u_{ij}]$ *where*

$$u_{ij} = u_{ij}(m) = \sum_{r=m-j}^{m} (-1)^{m-r} \left[{m \atop r}\right] \binom{r}{m-j} (m-1-i)^{r-(m-j)}$$

for $0 \le i, j < m$ *and where* 0^0 *is taken to be* 1.

Proof: We shall show that $\sum_{k=0}^{m-1} u_{ik}v_{kj} = m!\delta_{ij}$. We start with the standard trick of interchanging the order of summation:

$$\sum_{k=0}^{m-1} u_{ik}v_{kj} = \sum_{k=0}^{m-1} u_{ik} \sum_{r=0}^{j} (-1)^r \binom{m+1}{r}(j+1-r)^{m-k}$$
$$= \sum_{r=0}^{j} (-1)^r \binom{m+1}{r} \sum_{k=0}^{m-1} u_{ik}(j+1-r)^{m-k}.$$

Here we rewrite the inner sum as follows (note that $\left[{m \atop 0}\right] = 0$ in the second line and the interchange trick is used again in the third line):

$$\begin{aligned}
&\sum_{k=0}^{m-1} u_{ik}(j+1-r)^{m-k}\\
&\quad= \sum_{k=1}^{m} u_{i,m-k}(j+1-r)^{k}\\
&\quad= \sum_{k=0}^{m}\sum_{s=k}^{m} (-1)^{m-s}\left[{m \atop s}\right]\binom{s}{k}(m-1-i)^{s-k}(j+1-r)^{k}\\
&\quad= \sum_{s=0}^{m} (-1)^{m-s}\left[{m \atop s}\right]\sum_{k=0}^{s}\binom{s}{k}(m-1-i)^{s-k}(j+1-r)^{k}\\
&\quad= \sum_{s=0}^{m} (-1)^{m-s}\left[{m \atop s}\right](m-1-i+j+1-r)^{s} \qquad \text{[by the binomial theorem]}\\
&\quad= (m-i+j-r)^{\underline{m}} \qquad \text{[by (13)]}\\
&\quad= m!\binom{m-i+j-r}{m}.
\end{aligned}$$

Therefore,

$$\sum_{k=0}^{m-1} u_{ik}v_{kj} = m!\sum_{r=0}^{j} (-1)^r \binom{m+1}{r}\binom{m-i+j-r}{m}.$$

Note that $\binom{m-i+j-r}{m} = 0$ if $0 \le m-i+j-r < m$, i.e., $j-i < r \le m-i+j$. If $0 \le j < i < m$, every term in the last equation is 0; if $0 \le i = j < m$, only the $r = 0$ term is nonzero (it is $m!$); if $0 \le i < j < m$, we may add zero terms to get

$$\sum_{r=0}^{m+1} (-1)^r \binom{m+1}{r}\binom{m-i+j-r}{m} = \Delta^{m+1}\,(\text{polynomial in } r \text{ of degree } m) = 0.$$

Therefore,

$$\sum_{k=0}^{m-1} u_{ik}v_{kj} = m!\delta_{ij}. \qquad \blacksquare$$

CONFIRMATION OF THE CONJECTURES. The formula for **U** is complicated enough that it still takes some work to verify our conjectures. Conjectures 5, 6, and 8 are left as an exercises, and we turn to Conjecture 7, which claims that the top

row of $\mathbf{U}$ contains unsigned Stirling numbers of the first kind: For $j = 0, \ldots, m-1$,

$$u_{0j}(m) = \left[\begin{matrix} m \\ m-j \end{matrix}\right].$$

This conjecture neatly reduces to the two basic identities relating Stirling numbers of the first kind to factorial powers. By Theorem 4, for $n = 1, \ldots, m$,

$$u_{0,m-n}(m) = \sum_{r=n}^{m} (-1)^{m-r} \left[\begin{matrix} m \\ r \end{matrix}\right] \binom{r}{n} (m-1)^{r-n}.$$

Thus, Conjecture 7 is equivalent to the identity

$$\sum_{r=n}^{m} (-1)^{m-r} \left[\begin{matrix} m \\ r \end{matrix}\right] \binom{r}{n} (m-1)^{r-n} = \left[\begin{matrix} m \\ n \end{matrix}\right]. \tag{14}$$

Fix $m \geq n$, and let a_n denote the left side of (14). Switching the order of summation, we find that the generating function of $\langle a_n \rangle$ is

$$\begin{aligned}
\sum_{n \geq 0} a_n x^n &= \sum_{r=0}^{m} (-1)^{m-r} \left[\begin{matrix} m \\ r \end{matrix}\right] \sum_{n=0}^{r} \binom{r}{n} x^n (m-1)^{r-n} \\
&= \sum_{r=0}^{m} (-1)^{m-r} \left[\begin{matrix} m \\ r \end{matrix}\right] (x+m-1)^r \quad \text{[by the binomial theorem]} \\
&= (x+m-1)^{\underline{m}} \quad \text{[by (13)]} \\
&= (x+m-1)(x+m-2)\cdots(x) \\
&= x^{\overline{m}},
\end{aligned}$$

which is the generating function of $\left\langle \left[\begin{smallmatrix} m \\ n \end{smallmatrix}\right] \right\rangle$, by (12). Therefore, $a_n = \left[\begin{smallmatrix} m \\ n \end{smallmatrix}\right]$, i.e., identity (14) holds, so Conjecture 7 is true.

FURTHER CONSEQUENCES AND EXPLORATIONS. Many people are fascinated by combinatorial identities like (14), and there are many to be found in the context of the carries transition matrix. For example, the empirical recurrence identity for $\mathbf{V}$, (7), is indeed true, and provides a family of arrays satisfying the Eulerian recurrence (4). Other identities, including familiar ones, may be extracted from the matrix equations $\mathbf{\Pi U} = \mathbf{UD}$, $\mathbf{V\Pi} = \mathbf{DV}$, $\mathbf{UV} = m!\mathbf{I}$, and $\mathbf{VU} = m!\mathbf{I}$. Here is just one illustration: Set $i = 0$ and $j = m-1 > 0$ in $\Sigma_k v_{ik} u_{kj} = m! \delta_{ij}$ and get

$$\sum_{k=0}^{m-1} (-1)^k \left\langle \begin{matrix} m \\ k \end{matrix} \right\rangle \bigg/ \binom{m-1}{k} = 0.$$

Besides identities, we have seen geometric sequences, binomial coefficients, Eulerian numbers, and Stirling numbers of the first kind. What about other special numbers, like Stirling numbers of the second kind? Are they lurking nearby? Yes, indeed. Stirling numbers of the second kind crop up naturally in formulas for the factorial moments of the stationary probability distribution of $\mathbf{\Pi}$; alternatively, the nth factorial moment is exactly the generalized Bernoulli number $B_n^{(n-m)}$ (see [**5**, chapter 15]). A different sort of result is that the stationary probabilities are asymptotically normally distributed [**5**, pp. 150–154]. I hope some readers are inspired to discover other interesting connections.

Going beyond these kinds of propositions, we may put the matrix $\mathbf{\Pi}$ in a larger context: It is the $x = 1$ case of the matrix $[\pi_{ij} x^j]$, which plays a central role in the analysis of the asymptotic prime-power divisibility of multinomial coefficients [**9**].

Our tour has revealed a combinatorial richness hidden in the matrix **Π**. But it leaves unanswered the question, *Why* are all these combinatorially significant relationships connected with the carries matrix?

ACKNOWLEDGMENTS. I thank Jennifer Galovich and Paul Fjelstad for pointing out the connections with Eulerian numbers, and I thank Paul Fjelstad, Stephen Hilding, Ron Rietz, and Herbert Wilf for their comments on earlier drafts of this paper.

REFERENCES

1. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, D.C., 1964.
2. J. Buhler, D. Eisenbud, R. Graham, and C. Wright, Juggling drops and descents, *Amer. Math. Monthly* 101 (1994), 507–519.
3. D. Callan, M. P. Eisner, and B. Doran, A Pascal-like triangle, Problem #498 & solutions, *Coll. Math. J.* 25 (1994), 161–162.
4. L. Carlitz, Eulerian numbers and polynomials, *Math. Mag.* 32 (1959), 247–260.
5. F. N. David and D. E. Barton, *Combinatorial Chance*, Hafner, New York, 1962.
6. L. Euler, *Institutiones calculi differentialis*, Imperial Academy of Sciences, St. Petersburg, 1755.
7. L. Euler, *Opera Omnia* (1) 10, Teubner, Leipzig and Berlin, 1913.
8. R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics*, second edition, Addison-Wesley, Reading, Mass., 1994.
9. J. M. Holte, Asymptotic prime-power divisibility of binomial, generalized binomial, and multinomial coefficients, to appear in *Trans. Amer. Math. Soc.*
10. C. Jordan, *Calculus of Finite Differences*, Chelsea, New York, 1950.
11. D. E Knuth, *The Art of Computer Programming, vol. 1, Fundamental Algorithms*, second edition, Addison-Wesley, Reading, Mass., 1973.
12. D. E. Knuth, *The Art of Computer Programming, vol. 2, Seminumerical Algorithms*, second edition, Addison-Wesley, Reading, Mass., 1981.
13. D. E. Knuth, *The Art of Computer Programming, vol. 3: Sorting and Searching*, Addison-Wesley, Reading, Mass., 1973.
14. H. K. Krishnapriyan, Eulerian polynomials and Faulhaber's result on sums of powers of integers, *Coll. Math. J.* 26 (1995), 118–123.
15. D. Neal, The series $\Sigma_{n=1}^{\infty} n^m x^n$ and a Pascal-like triangle, *Coll. Math. J.* 25 (1994), 99–101.
16. A. Tucker, *Applied Combinatorics*, second edition, Wiley, New York, 1984.
17. J. R. Weaver, Centrosymmetric (cross-symmetric) matrices, their basic properties, eigenvalues, and eigenvectors, *Amer. Math. Monthly* 92 (1985), 711–717.
18. H. S. Wilf, *generatingfunctionology*, second edition, Academic Press, Boston, 1994.

Department of Mathematics and Computer Science
Gustavus Adolphus College
St. Peter, MN 56082
holte@gac,edu

A New Look at Euler's Theorem for Polyhedra: A Comment

Walter Nef

In their interesting paper [6] the authors, Branko Grünbaum and Geoffrey Shephard mention (page 125/126) "*... the work of Nef [20–25], which our note parallels to some extent. Nef's definition of polyhedral sets differs from ours, but is equivalent to it; his approach to the Euler characteristic is the same as ours.*" Actually, this parallelism concerns a major part of the article and comes partly into conflict with my own work in the field. The main reason is the authors' inexplicable statement (p. 126) "*Nef [1978] defines as faces of a polyhedral set P any family of disjoint, relatively open convex sets that is a dissection of P. ... However, these 'faces' are, in general, not uniquely determined, and have only a limited geometric significance.*" In fact my definition is the one in [8, p. 6.2], or in [1, p. 98]. According to this definition the faces of a polyhedron turn out to be the relative interiors of the faces in the "intuitive" sense. They are uniquely determined and have a clear geometric significance. (Unfortunately, in [1] a typing error slipped in: On p. 98 in formula (06) U_0 should be replaced by $P \cap U_0$, and U by $P \cap U$).

In a thorough discussion, several further points would have to be critically looked at. I confine myself to two of them:

On page 122, Grünbaum and Shephard present their Theorem 2* concerning the Euler characteristic of (not necessarily bounded) closed convex polyhedra. They overlook that I have published the same result previously in [9, Satz 4, p. 68].

On page 117, the authors define the Euler characteristic $\chi(P)$ (in the same way as I have done in [10, Satz 2, pp. 44–45]): For a *cell* C (a nonempty relatively open convex polyhedron) we put $\chi(C) := (-1)^{\dim C}$, furthermore, $\chi(\emptyset) = 0$. If $P = \bigcup_{C \in \mathbf{C}} C$ represents P as a finite disjoint union of cells, then $\chi(P) := \sum_{C \in \mathbf{C}} \chi(C)$. The definition is followed by three theorems, the first of which asserts that $\chi(P)$ does not depend on the partition of P into cells. (For a proof of this Theorem see [10, Satz 2, pp. 44–45]). The second theorem states that $\chi(P) = 1$ if P is bounded, closed, and convex. The proofs are omitted "*since they follow the usual techniques.*" From the discovery of his theorem by Euler (in 1750) more than 200 years were needed to find a complete proof ([5, p. 134], [7, p. 94], [2]); for an extended discussion of the historical development see [6, pp. 122–127]. It is therefore not evident what is meant by "the usual techniques," at least in the case of the second theorem, which is an essential part of the theorem of Euler-Schläfli (Leonhard Euler [3] and [4], and the extension to higher dimensions in 1850 by Ludwig Schläfli [12]). In [11] I have published a particularly short and simple proof which I repeat here: First of all, every closed polyhedron P is the disjoint union of its faces (without *ext* P) [8, Satz 3, p. 6.7]. If P is closed and *convex*, the faces are cells [8, Satz 8, p. 7.12]. Therefore $\chi(P) = \sum_{i=0}^{n}(-1)^i f_i$, where $n = \dim P$ and f_i is the number of faces of dimension i (with $f_n = 1$ for the relative interior of P).

Now let P be *bounded*, closed, and convex, and let Φ denote the family of all faces F of P with $\dim F < n$. We choose an arbitrary point $z \in$ *relint* P. For each face $F \in \Phi$ we define $F^* := \{x = z + \lambda(F - z) \text{ with } 0 < \lambda < 1\}$, so F^* is the

union of all open line-segments joining z with a point of F. It is intuitively clear and is easy to prove that the F^* too are cells, that $\dim F^* = \dim F + 1$, and that all F, F^* ($F \in \Phi$), together with $\{z\}$ form a partition of P (Figure 1). Therefore

$$\sum_{i=0}^{n} (-1)^i f_i = \chi(P) = \sum_{F \in \Phi} (\chi(F) + \chi(F^*)) + \chi(\{z\}) =$$
$$= \chi(\{z\}) = (-1)^0 = 1,$$

which is the theorem of Euler-Schläfli.

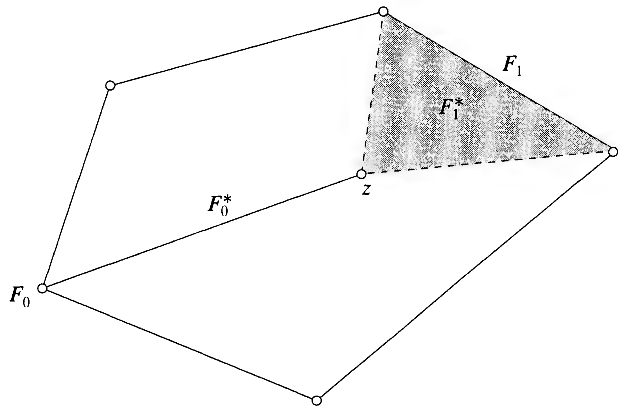


Figure 1

REFERENCES

1. H. Bieri and W. Nef, Elementary set operations with d-dimensional polyhedra, H. Noltemeier (ed): *Computational Geometry and its Applications. Lecture Notes in Computer Science* 333, Springer-Verlag, Berlin, Heidelberg, 1988, pp. 97–112.
2. H. Bruggesser and P. Mani, Shellable decompositions of cells and spheres, *Math. Scand.* 29 (1971), 197–205.
3. Leonhard Euler, Elementa doctrinae solidorum, *Opera omnia*, *series prima*, vol. 26, Orell Füssli, Zürich, 1953, pp. 71–93.
4. Leonhard Euler, Demonstratio nonnullarum insignium proprietatum quibus solida hedris planis inclusa sunt praedita, *Opera omnia*, *series prima*, vol. 26, Orell Füssli, Zürich, 1953, pp. 94–108.
5. B. Grünbaum, *Convex Polytopes*, John Wiley and Sons, London, 1967.
6. B. Grünbaum and G. C. Shephard, A new Look at Euler's Theorem for Polyhedra, *Amer. Math. Monthly* 101 (1994), 109–128.
7. P. McMullen and G. C. Shephard, Convex Polytopes and the Upper Bound Conjecture, *Mathematical Society Lecture Note Series*, vol. 3, Cambridge University Press, Cambridge, 1971.
8. W. Nef, *Beiträge zu Theorie der Polyeder mit Anwendungen in der Computergraphik*, Herbert Lang, Bern, 1978.
9. W. Nef, Zur Eulerschen Charakteristik allgemeiner, insbesondere konvexer Polyeder, *Resultate der Mathematik* 3 (1980), 64–69.
10. W. Nef, Zur Einführung der Eulerschen Charakteristik, *Monatshefte für Mathematik* 92 (1981), 41–46.
11. W. Nef, Ein einfacher Beweis des Satzes von Euler-Schläfli, *Elemente der Mathematik* 39 (1984), 1–6.
12. Ludwig Schläfli, Theorie der vielfachen Kontinuität, *Gesammelte Mathematische Abhandlungen*, Bd. 1, Birkhäuser, Basel, 1950, pp. 189–191.

Beundeweg 23
CH-3033 Wohlen b. Bern
Switzerland

NOTES

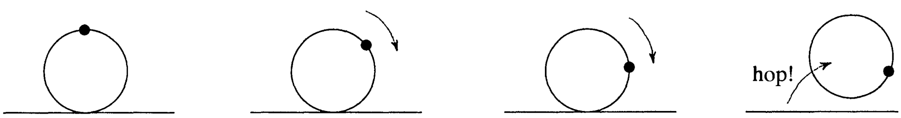
Edited by **Jimmie D. Lawson**

The Hopping Hoop

Tadashi F. Tokieda

'A weight is attached to a point of a rough weightless hoop, which then rolls in a vertical plane, starting near the position of unstable equilibrium. What happens, and is it intuitive?'

The problem just quoted is from Littlewood's delightful *Miscellany* [**1**, p.37]. As is often the case in phenomena involving a no-slip constraint ('rough' indicates that the hoop is to roll *without slipping*), what happens is rather unintuitive. Declares Littlewood: 'The hoop lifts off the ground when the radius vector to the weight becomes horizontal.'



Perhaps the most ingenuous approach to proving that the hoop indeed hops is to calculate the force that the hoop exerts against the floor at the point of contact, and to check that it changes to negative after the hoop has rolled $\pi/2$. The approach works, but hardly explains why the hoop should hop at all.

It is more pleasant to reason as follows. If the hoop is always kept in contact with the floor, then the weight (call it m) travels along a cycloid. Now imagine that, when m comes to a certain point P on the cycloid, the hoop suddenly disappears. Then m would continue to free-fall along a parabola tangent to the cycloid at P. If, however, the hoop fails to disappear (as it usually does), then

(1) m presses the hoop *down* as long as the imagined parabola at P departs *below* the cycloid;

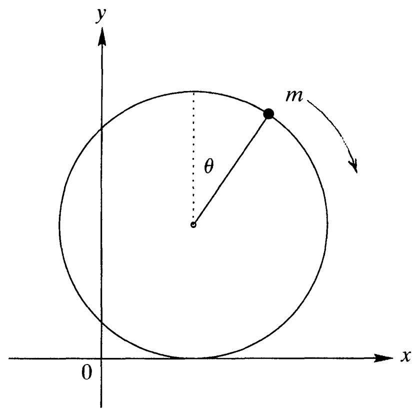
(2) m pulls the hoop *up*, and so the hoop hops, as soon as the imagined parabola at P departs *above* the cycloid.

By construction the parabola and the cycloid have the same zeroth and first terms in their Taylor series around P. Therefore departure below or above will be decided by an inequality between their *second* derivatives.

Let us determine the earliest P for which (2) occurs. We generalize the problem somewhat by taking the liberty of shoving m off the point of unstable equilibrium with initial velocity v_0 (in Littlewood's original formulation, $v_0 = 0$).

In coordinates as shown, with g and R denoting the gravitational acceleration and the radius of the hoop, the conservation of energy dictates

$$\frac{m}{2}\left(\dot{x}^2 + \dot{y}^2\right) + mgy = \frac{m}{2}v_0^2 + mg2R.$$



Along the cycloid

$$x(t) = R\theta(t) + R\sin\theta(t)$$
$$y(t) = R + R\cos\theta(t)$$

we have

$$\frac{m}{2}\left[\left(R\dot{\theta} + R\cos\theta\cdot\dot{\theta}\right)^2 + \left(-R\sin\theta\cdot\dot{\theta}\right)^2\right] + mg(R + R\cos\theta) = \frac{m}{2}v_0^2 + mg2R,$$

which unravels to

$$\dot{\theta}^2 = \frac{4gR\sin^2(\theta/2) + v_0^2}{4R^2\cos^2(\theta/2)}.$$

This relation enables us to express the derivatives of $y(t)$ in terms of θ:

$$\begin{aligned}\dot{y} &= -R\sin\theta\cdot\dot{\theta}\\ &= -\sin(\theta/2)\sqrt{4gR\sin^2(\theta/2) + v_0^2},\\ \ddot{y} &= -2g\sin^2(\theta/2) - \frac{v_0^2}{4R}.\end{aligned}$$

As remarked earlier, m pulls the hoop up, thereby making it hop, as soon as the second derivative of the parabola exceeds that of the cycloid; i.e. the hop occurs at minimal θ such that, $-g \geq \ddot{y}(\theta(t))$, or

$$\sin(\theta/2) \geq \frac{1}{\sqrt{2}}\left(1 - \frac{v_0^2}{4gR}\right)^{1/2}.$$

In particular, for $v_0 = 0$ the hoop hops at $\theta = \pi/2$, as claimed. We also observe that for $v_0 = \sqrt{4gR}$ the hoop 'glides' immediately without rolling. Naturally, this escape velocity ought to be larger than the escape velocity $\sqrt{2gR}$ for a circle: since the cycloid has smaller curvature than the circle does at the peak $\theta = 0$, it is harder to escape from the cycloid than from the circle.

The author thanks F. Almgren for collaborating on an experiment. We taped a battery on a hula-hoop and rolled it down the twelfth-floor hallway in Fine Hall; it actually hopped.

REFERENCES

1. J. E. Littlewood, *Littlewood's Miscellany*, Cambridge University Press, 1986.

Department of Mathematics
Princeton University
Princeton NJ 08544-1000
tokieda@math.princeton.edu

Principal Ideal Domains Are Almost Euclidean

John Greene

In most undergraduate level books on abstract algebra, it is shown that every Euclidean domain (ED) is a principal ideal domain (PID) and every principal ideal domain is a unique factorization domain (UFD). We thus have a set of implications: ED $\Rightarrow$ PID $\Rightarrow$ UFD. Most (but not all!) books mention that neither converse is true. But while it is very easy to show that $Z[x]$ is an example of a UFD that is not a PID, an example of a PID that is not a ED is harder to come by. In [**2**], Campoli gives an easy proof that $Z[\zeta]$ has the desired properties, where $\zeta = (-1 + \sqrt{-19})/2$, by showing that, in his words, $Z[\zeta]$ is "almost Euclidean." In this note, we show that Campoli's "almost Euclidean" condition is, in fact, equivalent to the PID condition.

Definition. An integral domain D is said to be **almost Euclidean** if there is a function $d\colon D \to Z^+ \cup \{0\}$ (called an almost Euclidean function) such that

1) $d(0) = 0$, $d(a) > 0$ if $a \neq 0$,
2) If $b \neq 0$, then $d(ab) \geq d(a)$ for all $a \in D$,
3) for any $a, b \in D$, if $b \neq 0$ then either
   i) $a = bq$ for some $q \in D$ or
   ii) $0 < d(ax + by) < d(b)$ for some $x, y \in D$.

Our functions d in this paper will satisfy the stronger condition (2′) that for all a, b in D, $d(ab) = d(a)d(b)$, from which (2) follows trivially.

Our main result is the following:

Theorem 1. *An integral domain D is a principal ideal domain if and only if it is almost Euclidean.*

Proof: Campoli [**2**] proved that if a ring is almost Euclidean, it is a PID. For completeness, we repeat the proof here. Let D be almost Euclidean, and let I be a nonzero ideal in D. Among the elements $x \in I$, let b be an element with a minimal positive value for $d(x)$. Given $a \in I$, for any $x, y \in D$, $ax + by$ is in I. By definition of b, it cannot be that $0 < d(ax + by) < d(b)$, so the second condition, $a = bq$ must hold for some $q \in D$. Thus, $I = (b)$.

Now suppose that D is a PID. Then D is a UFD, so we may define the function d as follows: Let $d(0) = 0$, and for any $a \neq 0$, if $a = \varepsilon p_1 p_2 \cdots p_n$, where ε is a unit and $p_1, \ldots, p_n$ are irreducibles, let $d(a) = 2^n$. Since $d(ab) = d(a)d(b)$, it is clear that d satisfies (1) and (2) of the definition. So let $a, b \in D$, with $b \neq 0$. Let $I = \{ax + by | x, y \in D\}$. Since I is an ideal in $D, I = \langle r \rangle$ for some $r \in D$ with $r \neq 0$. If $a = bq$ for some $q \in D$, then $I = \langle b \rangle$. Otherwise, $I \neq \langle b \rangle$. Since $b \in I, b = xr$ for some $x \in D$, so $d(b) \geq d(r)$. Since $I \neq \langle b \rangle$, x is not a unit. Thus, $d(x) > 1$, so $d(r) < d(b)$. If $r = x_0 a + y_0 b$, then we have $0 < d(r) < d(b)$, and condition (3) is satisfied by d. ∎

Examples of Euclidean domains in abstract algebra texts are almost always of the form $D = F[x]$, where F is a field or the ring of integers in $Q[\sqrt{d}\,]$ for various small integer values of d. In the latter case, these books introduce the norm of an element of this ring and use its absolute value as a Euclidean function. In general, if F is an algebraic number field (a finite extension of Q), then F can be viewed as a finite dimensional vector space over Q. If $a \in F$, then the map $T_a(x) = ax$ is obviously a Q-linear transformation from F to F. The norm of a, $N(a)$, is defined to be the determinant of this transformation. The norm has the following properties:

1) $N(ab) = N(a)N(b)$ for all $a, b \in F$,
2) $N(a) = 0$ if and only if $a = 0$,
3) if a is an algebraic integer, then $N(a) \in Z$,
4) an algebraic integer a is a unit if and only if $N(a) = \pm 1$,

Properties (1) and (2) are elementary properties of determinants, property (3) is mentioned in [**5**, p. 175], and property (4) is an easy consequence of (1), (2), and (3).

Theorem 2. *If D is the set of integers in an algebraic number field, and if D is a principal ideal domain, then the absolute value of the norm satisfies the conditions of an almost Euclidean function.*

Proof: The properties of the function d in Theorem 1 that were used in the proof were:

1) $d(ab) = d(a)d(b)$
2) if $a \in D$, then $d(a) = 1$ if and only if a is a unit.

Since the absolute value of the norm also has these properties, the proof follows as in Theorem 1. Thus, given $a, b \in D$, with $b \neq 0$, let $I = \{ax + by | x, y \in D\} = \langle r \rangle$. If $a = bq$ for some $q \in D$, then $I = \langle b \rangle$. Otherwise, $0 < |N(r)| < |N(b)|$, since $b = xr$ for some nonunit $x \in D$. ∎

If D is the ring of integers in some finite extension F of Q, we may now check if D is a principal ideal domain by checking whether or not D is almost Euclidean with respect to the absolute value of the norm. Thus, number fields are quite special. Another example of this is the following: In a number field, D is a UFD if and only if D is a PID [**6**, page 146]. Campoli [**2**] used the fact that $Z[\zeta]$ with $\zeta = (-1 + \sqrt{-19})/2$ is almost Euclidean to show that this ring is a PID. His techniques can be easily extended tc show that this remains true if -19 is replaced by -43 or -163. In fact, with a little work it is possible to prove the famous result

[1, p. 137]: The ring of integers in $Q(\sqrt{1-4d})$ where $d > 0$ is a PID if and only if the polynomial $x^2 + x + d$ is prime for all integers x with $0 \leq x \leq d - 2$.

One final comment: If D is an almost Euclidean subring of a number field, Theorem 2 tells us that we may use the absolute value of the norm as a near Euclidean function. Suppose that D is actually Euclidean. Will the absolute value of the norm serve as a Euclidean function? It is interesting to note that Hardy and Wright [4, p. 212] define a Euclidean domain not in the usual way but explicitly using the norm as the Euclidean function. However, the answer to the question is that the norm may not work. In fact, it was shown in [3] that $Z[\zeta]$, with $\zeta = (1 + \sqrt{69})/2$ is an example of a ring which is Euclidean, but not with respect to the absolute value of the norm.

ACKNOWLEDGMENTS. I would like to thank Joe Gallian and the reviewer for many helpful comments, and thank the members of the usenet newsgroup sci.math, especially Henry Cohn, for the reference to a ring that is Euclidean but not norm-Euclidean.

REFERENCES

1. Paulo Ribenboim, The Book of Prime Number Records, Springer-Verlag, New York, 1988.
2. Oscar Campoli, A Principal Ideal Domain That Is Not a Euclidean Domain, *American Mathematical Monthly*, **95** (1988) 868–871.
3. David Clark, A quadratic field which is Euclidean but not norm-Euclidean, *Manuscripta Math.* **83** (1994), no. 3-4, 327–330.
4. G. Hardy and E. Wright, *An Introduction to the Theory of Numbers*, 5th Edition, Oxford University Press, Oxford.
5. K. Ireland and M. Rosen, *A Classical Introduction to Modern Number Theory*, Springer-Verlag, New York, 1992.
6. Edwin Weiss, *Algebraic Number Theory*, McGraw-Hill, New York, 1963.

A Colorful Determinantal Identity, a Conjecture of Rota, and Latin Squares

Shmuel Onn

1. Rota's Colorful Conjecture and the Latin Square Conjecture. The following conjecture in combinatorial linear algebra is due to Gian-Carlo Rota.

Rota's Colorful Conjecture. Let ${}^1W, \ldots, {}^nW$ be bases of an n-dimensional vector space. Then their multiset union can be repartitioned into bases ${}^1U, \ldots, {}^nU$ such that $|{}^iU \cap {}^jW| = 1$ for all i, j.

Regarding the vectors in each ${}^iW$ as *colored* in color i, the newly sought bases are *colorful*, namely contain one vector of each color. So Rota's Colorful Conjecture is that any n colored bases of an n-dimensional vector space can be repartitioned into n colorful bases.

A *Latin square* of order n is an n by n matrix $L = (L_i^j)$ in which each row and each column is a permutation of $\{1, \ldots, n\}$. More precisely, there are permutations $\sigma_1, \ldots, \sigma_n$ and $\pi_1, \ldots, \pi_n$ such that $L_i^j = \sigma_i(j) = \pi_j(i)$ for all i, j. The *sign* of the Latin square is defined as the product of all signs of its row and column

permutations

$$\operatorname{sgn}(L) = \prod_{i=1}^{n} \operatorname{sgn}(\sigma_i) \cdot \operatorname{sgn}(\pi_i).$$

A Latin square is *even* if its sign is positive, and *odd* otherwise. Let $l(n)$ be the number of even Latin squares of order n minus the number of odd ones. It is easy to see that $l(n) = 0$ for all odd n. The following conjecture is due to Noga Alon and Michael Tarsi (cf. [1]).

Latin Square Conjecture. The number of even Latin squares of order n minus the number of odd ones satisfies $l(n) \neq 0$ for all even n.

This conjecture has been recently proved by Drisko [4] for all $n = p + 1$ where p is a prime. The exact values of $l(n)$ are known only for $n \leq 8$ [7]. It is plausible that $l(n)$ is in fact always nonnegative.

In this note we establish an identity, the *Colorful Determinantal Identity*, which links the two conjectures. It shows that for any n, the Latin Square Conjecture implies Rota's Colorful Conjecture. In particular, Rota's Colorful Conjecture is true for any $n = p + 1$ where p is a prime. To compactly express the identity, let S_n be the symmetric group of permutations on $\{1, \ldots, n\}$ and denote by $\mathscr{S}^n$ the collection of n-tuples $\rho = (\rho_1, \ldots, \rho_n)$ of permutations. For $\rho \in \mathscr{S}^n$ let $\operatorname{sgn}(\rho) = \prod_{i=1}^{n} \operatorname{sgn}(\rho_i)$. For a matrix W let W^j be its j-th column vector. The proof of the following theorem is given in Section 2.

Theorem 1 (Colorful Determinantal Identity). *Let ${}^1W, \ldots, {}^nW$ be n square matrices of order n over an arbitrary field. Then*

$$\sum_{\rho \in \mathscr{S}^n} \operatorname{sgn}(\rho) \prod_{i=1}^{n} \det({}^1W^{\rho_1(i)}, \ldots, {}^nW^{\rho_n(i)}) = l(n) \cdot \prod_{j=1}^{n} \det({}^jW).$$

Note that for each $\rho \in \mathscr{S}^n$, the tuples $({}^1W^{\rho_1(i)}, \ldots, {}^nW^{\rho_n(i)})$, $i = 1, \ldots, n$, which appear in the left hand side of the identity, give a colorful repartition of the multiset of column vectors of the ${}^jW$.

Corollary 1. *For any even n the Latin Square Conjecture implies Rota's Colorful Conjecture over any field of characteristic which does not divide $l(n)$ (in particular characteristic zero).*

Proof of Corollary 1 from Theorem 1. Suppose that $l(n) \neq 0$ and let ${}^1W, \ldots, {}^nW$ be given bases, i.e. nonsingular matrices of order n over a field satisfying the hypothesis. Then the right hand side of the Colorful Determinantal Identity is nonzero. Therefore, on the left hand side, there must exist a nonzero summand

$$\prod_{i=1}^{n} \det({}^1W^{\rho_1(i)}, \ldots, {}^nW^{\rho_n(i)}), \text{ and so the sets } {}^iU = \{{}^1W^{\rho_1(i)}, \ldots, {}^nW^{\rho_n(i)}\}, \quad i = 1, \ldots, n,$$

give the desired repartition into colorful bases. ■

Corollary 1 had been independently obtained by Huang and Rota, but their derivation is quite complex and indirect and involves an intermediate conjecture on Rota on a certain straightening coefficient in the so-called *supersymmetric bracket algebra* [6].

Rota's Colorful Conjecture has a natural generalization to *matroids* **[6]** which had been verified only for $n = 3$ **[3]**. Another generalization of Rota's conjecture is a conjecture of Jeff Kahn (cf. **[6]**) that concerns n^2 bases, for which we have derived a determinantal identity which is reminiscent of Theorem 1. The special case of Kahn's conjecture in which the vectors are in *general position* is known as the *Dinitz problem* and was recently settled affirmatively by Galvin **[5]**. We refer the reader to **[2]** for some related algorithmic problems and a discussion of their computational complexity.

2. The Colorful Determinantal Identity

Theorem 1 (Colorful Determinantal Identity). *Let ${}^1W, \ldots, {}^nW$ be n square matrices of order n over an arbitrary field. Then*

$$\sum_{\rho \in \mathscr{S}^n} \operatorname{sgn}(\rho) \prod_{i=1}^{n} \det({}^1W^{\rho_1(i)}, \ldots, {}^nW^{\rho_n(i)}) = l(n) \cdot \prod_{j=1}^{n} \det({}^jW).$$

Proof: For a matrix W let W^j be its j-th column vector as before, and let W_i denote its i-th row vector and W_i^j its (i, j)-th entry. Given n square matrices ${}^1W, \ldots, {}^nW$, of size n, define the following polynomial in their entries:

$$\Delta = \sum_{\sigma, \rho \in \mathscr{S}^n} \operatorname{sgn}(\sigma) \operatorname{sgn}(\rho) \prod_{i,j=1}^{n} {}^jW_{\sigma_i(j)}^{\rho_j(i)}.$$

For each ρ and σ in $\mathscr{S}^n$ define

$$\begin{aligned} \Delta^\rho &= \sum_{\sigma \in \mathscr{S}^n} \operatorname{sgn}(\sigma) \prod_{i,j=1}^{n} {}^jW_{\sigma_i(j)}^{\rho_j(i)} = \prod_{i=1}^{n} \sum_{\sigma_i \in S_n} \operatorname{sgn}(\sigma_i) \prod_{j=1}^{n} {}^jW_{\sigma_i(j)}^{\rho_j(i)} \\ &= \prod_{i=1}^{n} \det({}^1W^{\rho_1(i)}, \ldots, {}^nW^{\rho_n(i)}), \\ \Delta_\sigma &= \sum_{\rho \in \mathscr{S}^n} \operatorname{sgn}(\rho) \prod_{i,j=1}^{n} {}^jW_{\sigma_i(j)}^{\rho_j(i)} = \prod_{j=1}^{n} \sum_{\rho_j \in S_n} \operatorname{sgn}(\rho_j) \prod_{i=1}^{n} {}^jW_{\sigma_i(j)}^{\rho_j(i)} \\ &= \prod_{j=1}^{n} \det\left({}^jW_{\sigma_1(j)}, \ldots, {}^jW_{\sigma_n(j)}\right). \end{aligned}$$

Now Δ_σ is nonzero only for $\sigma = (\sigma_1, \ldots, \sigma_n) \in \mathscr{S}^n$ for which there exists an element $\pi = (\pi_1, \ldots, \pi_n)$ in $\mathscr{S}^n$ such that $\sigma_i(j) = \pi_j(i)$ for all i, j, in which case

$$\begin{aligned} \Delta_\sigma &= \prod_{j=1}^{n} \det\left({}^jW_{\pi_j(1)}, \ldots, {}^jW_{\pi_j(n)}\right) = \operatorname{sgn}(\pi) \prod_{j=1}^{n} \det\left({}^jW_1, \ldots, {}^jW_n\right) \\ &= \operatorname{sgn}(\pi) \prod_{j=1}^{n} \det({}^jW). \end{aligned}$$

Each $\sigma \in \mathscr{S}^n$ for which such $\pi \in \mathscr{S}^n$ exists gives a Latin square L via $L_i^j = \sigma_i(j) = \pi_j(i)$, whose sign is given by $\operatorname{sgn}(L) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\pi)$. Denote by $\mathscr{L}$ the set of all Latin squares of order n. Then

$$\begin{aligned} \sum_{\rho \in \mathscr{S}^n} \operatorname{sgn}(\rho) \Delta^\rho = \Delta = \sum_{\sigma \in \mathscr{S}^n} \operatorname{sgn}(\sigma) \Delta_\sigma &= \sum_{L \in \mathscr{L}} \operatorname{sgn}(L) \prod_{j=1}^{n} \det({}^jW) \\ &= l(n) \prod_{j=1}^{n} \det({}^jW), \end{aligned}$$

which is precisely the Colorful Determinantal Identity. ■

ACKNOWLEDGMENT. The author thanks Noga Alon and Ron Holzman for some related discussions, and acknowledges support from the Alexander von Humboldt Stiftung and the Fund for Promotion of Research at the Technion.

REFERENCES

1. N. Alon and M. Tarsi, Colorings and orientations of graphs, *Combinatorica*, 12:125–134, 1992.
2. I. Bárány and S. Onn, Colorful linear programming, *Lecture Notes in Computer Science*, 1084:1–15, 1996, Springer-Verlag.
3. W. Chan, An exchange property of matroid, *Discr. Math.*, 146:299–302, 1995.
4. A. A. Drisko, On the number of even and odd Latin squares of order $p + 1$, *Adv. Math.*, to appear.
5. F. Galvin, The list chromatic index of a bipartite multigraph, *J. Combin. Theory Ser. B*, 63:153–158, 1995.
6. R. Huang and G. C. Rota, On the relations of various conjectures on Latin squares and straightening coefficients, *Discr. Math.*, 128:225–236, 1994.
7. J. C. M. Janssen, On even and odd Latin squares, *J. Combin. Theory Ser. A*, 69:173–181, 1995.

Department of Operations Research
Technion–Israel Institute of Technology
32000 Haifa, Israel
onn@ie.technion.ac.il

Very Semisimple Modules

W. K. Nicholson

Throughout this note R will denote an associative ring with unity. A left module M over R is called *semisimple* if it is a (possibly infinite) sum of simple submodules, equivalently [**1**, p.437], if every submodule of M is a direct summand of M. The theory of these modules is well known and is basic to the study of noncommutative rings. We call a module M *very semisimple* if every principal submodule Rm, $0 \neq m \in M$, is simple. It is clear that every such module is semisimple; in this note we characterize when the converse is true.

Every simple module is very semisimple, as is every vector space over a field (or a division ring). If p is a prime in the ring $\mathbb{Z}$ of integers, and if $\mathbb{Z}_p = \mathbb{Z}/p\mathbb{Z}$ denotes the ring of integers modulo p, the direct sum $\mathbb{Z}_p^{(I)}$ of $|I|$ copies of $\mathbb{Z}_p$ is very semisimple as a $\mathbb{Z}$-module because each nonzero element has order p.

On the other hand, $M = \mathbb{Z}_2 \oplus \mathbb{Z}_3$ is semisimple as a $\mathbb{Z}$-module but it is not very semisimple because $M = \mathbb{Z}m$ is not simple where $m = (1 + 2\mathbb{Z}, 1 + 3\mathbb{Z})$. As the following proposition shows, the reason is that $\mathbb{Z}_2$ and $\mathbb{Z}_3$ are not isomorphic. Recall that a semisimple module M is called *homogeneous* if any two simple submodules of M are isomorphic. We are going to show that every very semisimple module is homogeneous, and the following characterization will be needed.

Lemma 1. *A semisimple left R-module M is very semisimple if and only if $R(m_1 + m_2)$ is simple whenever Rm_1 and Rm_2 are simple where m_1 and m_2 are in M and $m_1 + m_2 \neq 0$.*

Proof: If $0 \neq m \in M$ let $m \in Rx_1 \oplus \cdots \oplus Rx_k$ where each Rx_i is simple [**1**, p. 437]. Write $m = m_1 + \cdots + m_k$ where $m_i \in Rx_i$ for each i. We may assume that

each $m_i \neq 0$, so $Rm_i = Rx_i$ is simple. Hence $R(m_1 + m_2)$ is simple by hypothesis ($m_1 + m_2 \neq 0$ because $Rm_1 \oplus Rm_2$ is direct). Again, $(m_1 + m_2) + m_3 \neq 0$ so $R(m_1 + m_2 + m_3)$ is simple. Continue to conclude that $Rm = R(m_1 + \cdots + m_k)$ is simple. ■

Proposition. *Every very semisimple module is homogeneous.*

Proof: If Rm and Rm' are simple where m and m' are in M, we must show that $Rm \cong Rm'$. We may assume that $Rm \neq Rm'$. Define α: $R(m + m') \to Rm$ and β: $R(m + m') \to Rm'$ by $\alpha[r(m + m')] = rm$ and $\beta[r(m + m')] = rm'$. These are well defined because $Rm \oplus Rm'$ is direct ($Rm \neq Rm'$), so both are onto. But $R(m + m')$ is simple by Lemma 1, so α and β are both isomorphisms, as required. ■

We hasten to note that not every homogeneous semisimple module is very semisimple, and the theorem below will tell us exactly which ones are.

If X is a subset of a left R-module, denote its annihilator by $l(X) = \{r \in R | rx = 0 \text{ for all } x \in X\}$. We abbreviate $l(\{m\}) = l(m)$. For convenience, we write $A \lhd R$ to mean that A is a two-sided ideal of the ring R. Our general characterization of very semisimple modules depends on the following fact about simple modules.

Lemma 2. *The following conditions are equivalent for a simple left R-module M*:

(1) $l(m) \lhd R$ *for every* $m \in M$.
(2) $l(m) \lhd R$ *for some* $0 \neq m \in M$.
(3) $M \cong R/A$ *where* $A \lhd R$ *and* A *is maximal as a left ideal of* R.
(4) $l(m) = l(m')$ *for all* $m \neq 0$ *and* $m' \neq 0$ *in* M.

Proof: Clearly (1) $\Rightarrow$ (2), and (2) $\Rightarrow$ (3) using $A = l(m)$. Assume (3), and let σ $M \to R/A$ be an R-isomorphism where A is as in (3). Given $0 \neq m \in M$, write $\sigma m = b + A$, $b \in R$. Then $l(m) = l(\sigma m) = \{r | rb \in A\}$. Because A is an ideal, we have $A \subseteq l(m)$, and so $A = l(m)$ because A is a maximal left ideal. This proves that (3) $\Rightarrow$ (4).

Finally, if (4) holds, let $m \in M$; we must show that $l(m) \lhd R$. This is clear if $m = 0$. If $m \neq 0$ write $A = l(m)$ so that A is a left ideal of R. We must show that $Ar \subseteq A$ for all $r \in R$. This is clear if $r \in A$; if $r \notin A$ write $m' = rm \neq 0$. Then $A = l(m')$ by (4) so $0 = Am' = Arm$. Thus $Ar \subseteq A$ and (1) follows. ■

Lemma 3. *The following conditions are equivalent for a nonzero, homogeneous, semisimple left R-module M*:

(1) $l(m) = l(m')$ *for all* $m \neq 0$ *and* $m' \neq 0$ *in* M.
(2) $l(m) \lhd R$ *for any* $m \in M$ *with* Rm *simple.*
(3) $l(m) \lhd R$ *for some* $m \in M$ *with* Rm *simple.*
(4) $l(M)$ *is maximal as a left ideal of* R.
(5) $M \cong (R/A)^{(I)}$ *for some set* I *and* $A \lhd R$ *with* A *maximal as a left ideal.*

In this case M is very semisimple, and if $A = l(M)$ *then* $\Delta = R/A$ *is a division ring and* $end(_R M) = end(_\Delta M)$ *where M is a left* Δ*-space via* $(r + A)m = rm$ *for all* $r \in R$ *and* $m \in M$.

Proof:

(1) $\Rightarrow$ (2). Let Rm be simple, $m \in M$. Then $l(m) = l(m')$ for all $0 \neq m' \in Rm$ by (1), so $l(m) \lhd R$ by Lemma 2.

(2) $\Rightarrow$ (3). This is clear.

(3) $\Rightarrow$ (4). By (3) fix $m_0 \in M$ with Rm_0 simple and $l(m_0) \lhd R$. Write $A = l(m_0)$.

Claim. $l(m') = A$ for all $m' \in M$ with Rm' simple.

Proof: $Rm_0 \cong Rm'$ because M is homogeneous. If $\sigma\colon Rm_0 \to Rm'$ is an isomorphism, then $l(\sigma m_0) = l(m_0) = A \lhd R$. As $\sigma m_0 \in Rm'$ Lemma 2 (applied to the simple module Rm') shows that $l(\sigma m_0) = l(m')$. This proves the Claim.

Now let $0 \neq m \in M$. Then $m \in Rx_1 \oplus \cdots \oplus Rx_k$ where each Rx_i is simple, so assume $m = m_1 + \cdots + m_k$ where $Rm_i = Rx_i$ for each i. Thus $Am_i = 0$ for each i by the Claim, so $Am = 0$. This means $A \subseteq l(M)$, and so $A = l(M)$ because A is a maximal left ideal. This proves (4).

(4) $\Rightarrow$ (5). Take $A = l(M)$. Then $A \subseteq l(m) \neq R$ for all $0 \neq m \in M$. Hence $A = l(m)$ by (4), so $Rm \cong R/A$. Now (5) follows because M is a direct sum of simple submodules.

(5) $\Rightarrow$ (1). Let $0 \neq m \in M$; we must show that Rm is simple. By (5) we may assume that $M = (R/A)^{(I)}$, so $Am = 0$ because A is an ideal. Thus $A \subseteq l(m)$, so $A = l(m)$ because A is a maximal left ideal. This proves (1).

Finally, note that $Rm \cong R/A$ is simple in (5) $\Rightarrow$ (1), so M is very semisimple. If $A = l(M)$ and $\Delta = R/A$, then Δ is a division ring by (4) because 0 is a maximal left ideal. Since $(r + A)m = rm$ is a well defined Δ-action on M, the last statement follows. ■

Note that the last statement in Lemma 3 actually holds for any homogeneous, semisimple module except that the ring Δ could be any left primitive ring rather than a division ring.

Theorem. *Let M be a nonzero homogeneous, semisimple left R-module. Then M is very semisimple if and only if either M is simple or M is non-simple and satisfies the conditions in Lemma* 3.

Proof: By Lemma 3, it remains to show that if M is very semisimple but not simple then $l(m) \lhd R$ for every $m \in M$ with Rm simple. By Lemma 2 it suffices to show that $l(m') = l(m)$ for all $0 \neq m' \in Rm$. Since M is not simple choose $k \in M$ with $k \notin Rm$. Then Rk is simple because M is very semisimple, so $Rm \oplus Rk$ is direct. This implies that $l(m + k) \subseteq l(m) \cap l(k)$. Since $R(m + k)$ is simple it follows that $l(m) = l(m + k) = l(k)$. Similarly $l(m') = l(k)$. ■

Corollary 1. *The following conditions on a ring R are equivalent*:

(1) *Every maximal left ideal of R is two-sided.*
(2) *Every homogeneous, semisimple left R-module is very semisimple.*
(3) *Every homogeneous, semisimple left R-module is length* 2 *is very semisimple.*

Proof: Lemma 3 gives (1) $\Rightarrow$ (2) and (2) $\Rightarrow$ (3) is clear. If A is a maximal left ideal of R, then $M = R/A \oplus R/A$ is very semisimple by (3). Let $r \in R$ and consider $m = (1 + A, r + A)$ in M. It suffices to show that $Am = 0$ (then $0 = (A, ar + A)$ for all $a \in A$). But if $Am \neq 0$ then $Am = Rm$ because Rm is simple (M is very

semisimple). Consequently $m = am$ for some $a \in A$. This means $(1 + A, r + A) = (A, ar + A)$, so $1 \in A$, a contradiction. Hence (3) $\Rightarrow$ (1). ∎

The conditions in Corollary 1 are satisfied for every commutative ring and for every local ring (a ring is called local if the Jacobson radical is the only maximal left (or right) ideal). They also hold for the ring $R = \begin{bmatrix} \Delta & \Delta \\ 0 & \Delta \end{bmatrix}$ of upper triangular matrices over a division ring Δ. In fact the only maximal left ideals of R are $\begin{bmatrix} \Delta & \Delta \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & \Delta \\ 0 & \Delta \end{bmatrix}$, and both are two-sided.

If M is any module and K is any simple submodule, the sum of all submodules of M that are isomorphic to K is a submodule called the *homogeneous component* of M generated by K. It is well known that the *socle* of M (that is the sum of *all* simple submodules of M) is the direct sum of the various homogeneous components. If $A \lhd R$, write $r_M(A) = \{m \in M | Am = 0\}$.

Corollary 2. *Let $H \neq 0$ be a non-simple homogeneous component of a left R-module M. The following are equivalent:*

(1) *H is very semisimple.*
(2) *$H = r_M(A)$ for some $A \lhd R$ such that A is maximal as a left ideal of R.*

In this case, $A = l(H)$ and $Rm \cong R/l(H)$ for all $0 \neq m \in H$.

Proof: (1) $\Rightarrow$ (2). Choose any $m_0 \in H$ with Rm_0 simple, and write $A = l(m_0)$. Then A is a maximal left ideal of R, and $A \lhd R$ by the Theorem. Moreover, (1) of Lemma 3 shows that $A = l(m)$ for all $0 \neq m \in H$, so $H \subseteq r_M(A)$. But if $0 \neq x \in r_M(A)$ then $A \subseteq l(x)$. Hence $A = l(x)$ by the maximality of A, and so $Rx \cong R/A \cong Rm$. Thus $x \in H$ because H is a homogeneous component of M.

(2) $\Rightarrow$ (1). If $H = r_M(A)$ as in (2), then $A \subseteq l(H)$. Thus $A \subseteq l(m) \neq R$ for all $0 \neq m \in H$, so $A = l(m)$ because A is a maximal left ideal. But then $Rm \cong R/l(m) = R/A$ is simple, proving (1).

Finally, we have $A \subseteq l(H) \neq R$, so $A = l(H)$ by the maximality of A. Thus $Rm \cong R/A = R/l(H)$ by the proof of (2) $\Rightarrow$ (1). ∎

REFERENCE

1. T. W. Hungerford, *Algebra*, Holt, Rinehart and Winston, New York, 1974.

Department of Mathematics and Statistics
University of Càlgary
Calgary, Alberta, T2N 1N4, CANADA
wknichol@acs.ucalgary.ca

PROBLEMS AND SOLUTIONS

Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**
with the collaboration of Paul T. Bateman, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Underwood Dudley, Michael A. Filaseta, Ira M. Gessel, Bart Goddard, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Robert Israel, Murray S. Klamkin, Daniel J. Kleitman, Fred Kochman, Frederick W. Luttmann, Frank B. Miles, M. J. Pelling, Richard Pfiefer, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before July 31, 1997; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.*

PROBLEMS

10571. *Proposed by Jean-Claude Evard and Hillel Gauchman, Eastern Illinois University, Charleston, IL.* Let $x_1, \dots, x_n$ be nonnegative real numbers and set $\tilde{x} = \min\{x_1, \dots, x_n\}$. Consider

$$A_\alpha = \left(\sum_{i=1}^n x_i\right)^\alpha - \sum_{i=1}^n x_i^\alpha - \left(n^{\alpha-1} - 1\right) \cdot \tilde{x}^{\alpha-1} \cdot \sum_{i=1}^n x_i$$

for $\alpha \in \mathbb{R}$. If α is a positive integer, consideration of the terms of $\left(\sum_{i=1}^n x_i\right)^\alpha$ shows that $A_\alpha \geq 0$. Show that $A_\alpha \geq 0$ for all $\alpha \in (-\infty, 1] \cup [2, \infty)$.

10572. *Proposed by Richard P. Stanley, Massachusetts Institute of Technology, Cambridge, MA.* Let $f(n)$ be the number of graphs (without loops or multiple edges) on the vertices $1, 2, \dots, n$ such that no path of length three has vertices i, j, k (in that order) with $i < j < k$. Let $g(n)$ be the total number of subspaces of an n-dimensional vector space over a 2-element field. Show that

$$\sum_{n\geq 0} f(n)\frac{x^n}{n!} = e^{-x}\sum_{n\geq 0} g(n)\frac{x^n}{n!}.$$

10573. *Proposed by Y.-F. S. Pétermann, University of Geneva, Geneva, Switzerland.* Find a continuous function $f : [0, \infty) \to [0, \infty)$ satisfying $f(0) = 0$ and the functional differential equation $f'(t) = 1/f(f(t))$ for $t > 0$, and show that no other such function exists.

10574. *Proposed by Vartan O. Choulakian, Université de Moncton, Moncton, N. B., Canada.* Let $\mathrm{Si}(x) = \int_0^x (\sin t/t)\,dt$ denote the sine integral function. Show that

(a) $\displaystyle\sum_{n=1}^{\infty}\left(\frac{\mathrm{Si}(n\pi)}{n}\right)^2 = \frac{\pi^2}{2}$.

(b) $\displaystyle\sum_{n=1}^{\infty}\frac{\mathrm{Si}(n\pi)}{n^3} = \pi^3\left(\frac{1}{8} - \frac{1}{18}\right)$.

(c) $\displaystyle\sum_{n=1}^{\infty}(-1)^n\frac{\mathrm{Si}(n\pi)}{n^3} = -\frac{\pi^3}{18}$.

10575. *Proposed by Xiaokang Yu, Penn State Altoona Campus, Altoona, PA.* Prove that

$$\sum_{l=0}^{n}(-1)^l\binom{n}{l}(2l)!\sum_{m=0}^{2l}\frac{(-1)^m}{m!}=\sum_{l=0}^{n}(-1)^l\binom{n}{l}2^{n-l}(n+l)!$$

for every nonnegative integer n.

10576. *Proposed by Donald E. Knuth, Stanford University, Stanford, CA.* Alice and Bill have identical decks of 52 cards. Alice shuffles her deck and deals the cards face up into 26 piles of two cards each. Bill does the same with his deck. If any one of Alice's top cards exactly matches any of Bill's, the matching cards are removed. Play continues until none of the cards on top of Alice's piles matches any of the cards on top of Bill's piles. What is the probability that all 52 pairs of cards will be matched?

10577. *Proposed by Mark Bowron and Stanley Rabinowitz, MathPro Press, Westford, MA.* It is well known that a maximum of 14 distinct sets are obtainable from one set in a topological space by repeatedly applying the operations of closure and complement in any order. Is there any bound on the number of sets that can be generated if we further allow arbitrary unions to be taken in addition to closures and complements?

SOLUTIONS

A Pentagonal Maximum Problem

6642 [1990, 857]. *Proposed by the editors.* Let λ be the maximum possible inradius of an arbitrary triangle lying in the closed set bounded by a regular pentagon of side-length one.
(**a**)* Determine λ up to an error of at most 10^{-3}.
(**b**)* Determine λ exactly.

Cf. 6477[1984, 588; 1986, 406; 1989, 945] and 6478 [1984, 588; 1990, 858].

Solution of (**a**) *by Richard Stong, Rice University, Houston, TX.* With the aid of a computer we show that

$$0.37550128 \le \lambda \le 0.37650126.$$

First we note that enlarging the triangle increases the inradius. Thus, if a vertex of the triangle lies in the interior of the pentagon, then we may increase the inradius by moving that vertex (along the angle-bisector, say) until it reaches some edge of the pentagon. Thus we may assume that all three vertices of the triangle lie on edges of the pentagon.

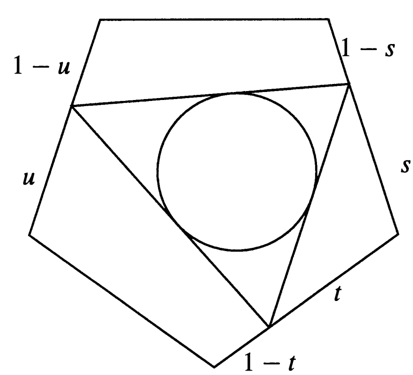


Figure 6642A

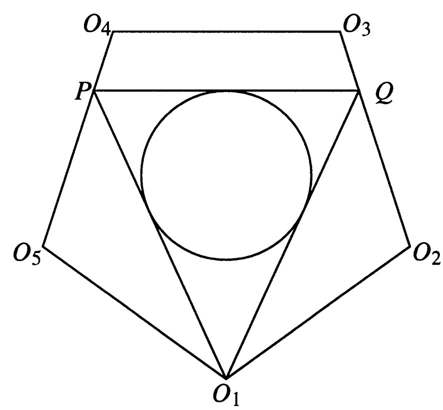


Figure 6642B

Further, if all three vertices of the triangle lie in one of the open half-planes bounded by some diagonal of the pentagon, then we can translate the triangle, with a motion perpendicular to that diagonal, into the interior of the pentagon and subsequently increase the radius. Therefore we may assume that the vertices of the triangle lie on three non-consecutive edges of the pentagon, as shown in Figure 6642A. This reduces the problem to considering a function defined for $(s, t, u) \in [0, 1] \times [0, 1] \times [0, 1]$. We need the following easy lemma.

Lemma. *Let ABC and DEF be two triangles in $\mathbb{R}^2$ and let $r(ABC)$ and $r(DEF)$ be their inradii. If each vertex of DEF is within ϵ of the corresponding vertex of ABC, then*

$$|r(ABC) - r(DEF)| \leq \epsilon.$$

Proof. By symmetry it is enough to show that $r(DEF) \geq r(ABC) - \epsilon$. We may assume that $r(ABC) > \epsilon$. Let G be the open circular disk centered at the incenter P of ABC with radius $r(ABC) - \epsilon$. Let L be the line tangent to the boundary of G and parallel to AB at a distance of ϵ from it. Since D and E are within ϵ of A and B respectively, both of these vertices lie on the opposite side of L from C. Hence the segment DE does not enter G. Similarly the segments EF and FD do not enter G. Further, a continuity argument involving the triangle with vertices $(1-\lambda)A + \lambda D$, $(1-\lambda)B + \lambda E$, $(1-\lambda)C + \lambda F$, where $0 \leq \lambda \leq 1$, shows that all of G, and in particular the point P, lies inside the triangle DEF. Thus $r(DEF) \geq r(ABC) - \epsilon$ and the lemma is proved.

We calculated the inradius of each of the $(501)^3$ triangles obtained by letting s, t, and u each run over the 501-element set $\{0.000, 0.002, 0.004, \ldots, 0.996, 0.998, 1.000\}$. A lengthy computer run found that the largest inradius of any of these $(501)^3$ triangles occurred when $s = 0.728$, $t = 1$, $u = 0.726$ (or when $s = 0.726$, $t = 1$, $u = 0.728$) and was $0.375501257\ldots$. (A value almost as large, namely $0.3742\ldots$ occurred when $s = t = 0.970$, $u = 0.500$). Any triangle with vertices on the three non-consecutive sides of the pentagon shown in Figure 6642A has each vertex within 0.001 of the corresponding vertices of one of these $(501)^3$ triangles. Thus by the lemma $\lambda \leq 0.37650126$. By taking $s = u = 0.726832$ and $t = 1$ we obtained the slightly larger inradius $0.375501286\ldots$, so that we can infer $\lambda \geq 0.37550128$.

Editorial comment. Part (**a**): The impetus for this problem was C. A. Rogers's solution of MONTHLY Problem 6478, which was based on the assumption that the inradius of any triangle lying in the closed set bounded by a regular pentagon of width 1 has an upper bound that is strictly less than 1/4. Since the regular pentagon of width 1 has sides of length $2\cot(2\pi/5)$, the maximum inradius of any triangle lying in the closed set bounded by a regular pentagon of width 1 is $2\lambda\cot(2\pi/5)$. Stong's results show that

$$0.2440 < 2\lambda\cot(2\pi/5) < 0.2447 < 1/4,$$

which shows that Rogers's assumption is justified.

Part (**b**): The exact value of λ is $2T^3/(1+T^2) = 0.37550128\ldots$, where $T = 0.64257343\ldots$ is the unique number in $(0, 1)$ such that

$$4T^3/(1-T^2)^2 = \tan(2\pi/5) = \sqrt{5+2\sqrt{5}}.$$

Alternatively, T is the unique solution in $(1/2, 1)$ of the algebraic equation

$$5T^{16} - 200T^{14} + 1036T^{12} - 1240T^{10} + 990T^8 - 440T^6 + 140T^4 - 40T^2 + 5 = 0.$$

The inradius $2T^3/(1+T^2)$ occurs for the triangle PO_1Q in the position shown in Figure 6642B, where $O_1O_2O_3O_4O_5$ is the given regular pentagon of side-length 1, PQ is parallel to O_3O_4, and angle $PO_1Q = \pi - 4\arctan T$. The two sides O_1P and O_1Q of the triangle PO_1Q have length $2T^2/(1-T^2) = 1.4062346\ldots$ and the third side PQ

has length $4T^2/(1+T^2) = 1.1688258\ldots$. It is not difficult (using the law of sines) to conclude further that the segments O_2Q and O_3P have length $0.72683402\ldots$.

Two solutions establishing the exact value of λ were received, one from Lou Hong-Wei of Ning-Bo University, Zhejiang Province, People's Republic of China, and the other from Li Wenzhi and Cheng Yiping of the University of Science and Technology, Hefei, People's Republic of China. In addition A. Tissier of Montfermeil, France obtained the same result under the restrictive assumption that the triangle in question is isosceles and its axis of symmetry coincides with one of the five axes of symmetry of the pentagon. Because of the considerable length and complicated nature of these solutions, we do not include a solution of (**b**) here.

A Special Sequence of Algebraic Integers

E 3461 [1991, 755]. *Proposed by David Callan, University of Wisconsin, Madison, WI.* Suppose r is a rational number but not an integer. It is known that $\tan(r\pi/2)$ is an algebraic number. (Cf. Ivan Niven, *Irrational Numbers*, Carus Mathematical Monographs No. 11, pp. 37-41.) Find the smallest positive integer k_r such that $k_r \tan(r\pi/2)$ is an algebraic integer.

Solution by Albert Nijenhuis, University of Pennsylvania (Emeritus), Philadelphia, PA, and University of Washington, Seattle, WA. If the denominator of r, in lowest terms, is a power of an odd prime p, then $k_r = p$; otherwise $k_r = 1$.

The cyclotomic polynomial $\Phi_n(z)$ of order n is the polynomial whose (simple) zeros are the primitive nth roots of unity $\left\{ e^{2\pi ik/n} : \gcd(n,k) = 1 \right\}$. Its degree is $\phi(n)$, the Euler totient function, and it is irreducible over the rationals. The nth roots of unity are the zeros of $z^n - 1$, and

$$z^n - 1 = \prod_{d|n} \Phi_d(z). \tag{1}$$

Let $P_n(t)$ be the polynomial $(1-it)^{\phi(n)}\Phi_n((1+it)/(1-it))$, for $n > 2$. In view of the relation $e^{i\theta} = (1+i\tan(\theta/2))/(1-i\tan(\theta/2))$, the zeros of $P_n(t)$ are the numbers $\tan(k\pi/n)$ such that $\gcd(k,n) = 1$. Since any factorization of $P_n(t)$ would yield a factorization of $\Phi_n(z)$, $P_n(t)$ is irreducible over the rationals.

When $n > 1$, the coefficients of z^j and $z^{\phi(n)-j}$ in $\Phi_n(z)$ are equal. This follows from (1) by induction on n and reflects the fact that if $\Phi_n(z_0) = 0$, then $\Phi_n(z_0^{-1}) = 0$. If $n > 2$, then $\phi(n)$ is even, and we define $\{a_j\}$ by $\Phi_n(z) = \sum_{j=0}^{\phi(n)/2} a_j[z^j + z^{\phi(n)-j}]$.

The constant term of $P_n(t)$ is $P_n(0) = \Phi_n(1)$, and since

$$\begin{aligned} P_n(t) &= (1-it)^{\phi(n)} \sum_{j=0}^{\phi(n)/2} a_j \left[\left(\frac{1+it}{1-it}\right)^j + \left(\frac{1+it}{1-it}\right)^{\phi(n)-j} \right] \\ &= \sum_{j=0}^{\phi(n)/2} a_j (1-it)^{\phi(n)/2}(1+it)^{\phi(n)/2} \left[\left(\frac{1-it}{1+it}\right)^{\phi(n)/2-j} + \left(\frac{1+it}{1-it}\right)^{\phi(n)/2-j} \right] \\ &= \sum_{j=0}^{\phi(n)/2} a_j (1+t^2)^j \left((1-it)^{\phi(n)-2j} + (1+it)^{\phi(n)-2j} \right) \end{aligned}$$

the leading coefficient of $P_n(t)$ is

$$\sum_{j=0}^{\phi(n)/2} a_j 2(-1)^{\phi(n)/2-j} = (-1)^{\phi(n)/2}\Phi_n(-1).$$

If n is a positive power of a prime p, then $\Phi_n(1) = p$. If n has distinct prime divisors, then $\Phi_n(1) = 1$. This is well known and follows from (1) by an induction argument applied to $\prod_{\{d:d|n,d>1\}} \Phi_d(1) = \lim_{z\to 1}(z^n - 1)/(z - 1) = n$.

Suppose $n > 2$. If n is odd, then $\Phi_n(-1) = 1$. If n is even, then

$$\prod_{\{d:d|n,d>2\}} \Phi_d(-1) = \lim_{z\to -1} \frac{z^n - 1}{z^2 - 1} = n/2.$$

Therefore (again by induction), $\Phi_n(-1) = 2$ if n is a power of 2, and $\Phi_n(-1) = p$ when n is twice a positive power of an odd prime. In all other cases, $\Phi_n(-1) = 1$.

When n is not a power of 2, either $\Phi_n(1) = 1$ or $\Phi_n(-1) = 1$, making $P_n(t)$ primitive (the greatest common divisior of its coefficients is 1). Otherwise, $\Phi_n(z) = z^{n/2} + 1$, and it is easy to see that $\frac{1}{2}P_n(t)$ is primitive. It follows that the leading coefficient of the primitive polynomial $P_n(t)$ (or $\frac{1}{2}P_n(t)$) equals ± 1 when n is odd, a power of 2, or twice a number with distinct prime divisors. In these cases, we set $K_n = 1$. When n is twice a positive power of an odd prime p, we set $K_n = p$. As a result, if $\gcd(k, n) = 1$, then $K_n \tan(k\pi/n)$ is a zero of a polynomial with leading coefficient ± 1, namely the polynomial $K_n^{\phi(n)-1} P_n(z/K_n)$ or half of this, and K_n is the minimal such integer.

Finally, if n is the denominator of the rational number $r/2$ when written in lowest terms, then we set $k_r = K_n$.

Editorial comment. Another method of proof, used by Robin Chapman and a referee, replaced the detailed study of the cyclotomic polynomials by corresponding properties of the algebraic numbers $1 - \zeta_n$, where ζ_n is a primitive nth root of unity. While the selected solution is more elementary, the discovery of the result and the organization of the proof can be simplified by the use of fairly well-known algebraic number theory.

Solved also by R. J. Chapman (U. K.), P. Čižek (Czech Republic), I. Kastanas, O. P. Lossers (The Netherlands), and the proposer.

Singular Values of a Classical Matrix

10312 [1993, 499]. *Proposed by Hongyuan Zha, IMA–University of Minnesota, Minneapolis, MN.* Let c and s be non-negative real numbers satisfying $c^2 + s^2 = 1$. Prove that, for $n > 1$, $s^{n-2}\sqrt{1+c}$ is the *second* smallest singular value of the n by n upper triangular matrix

$$T_n(c) = \operatorname{diag}(1, s, \dots, s^{n-1}) \begin{pmatrix} 1 & -c & -c & \cdots & -c \\ & 1 & -c & \cdots & -c \\ & & \ddots & \ddots & \vdots \\ & & & 1 & -c \\ & & & & 1 \end{pmatrix}.$$

Solution I by the proposer. Let $\sigma = s^{n-2}\sqrt{1+c}$. If $s = 0$ or $s = 1$, the result is obvious, so we assume $0 < s < 1$. Let $v^{\mathrm{T}} = \left(1/\sqrt{2}\right)(0, \dots, 0, 1, -1)$, and let $u^{\mathrm{T}} = \left(1/\sqrt{2(1+c)}\right)(0, \dots, 0, 1+c, -s)$; these are n-dimensional vectors of length 1. Since $T_n(c)v = \sigma u$ and $T_n(c)^{\mathrm{T}}u = \sigma v$, it follows that σ is a singular value of $T_n(c)$. With $\sigma_1 \geq \cdots \geq \sigma_n$ being the singular values of $T_n(c)$, the inequality $\sigma_n(T_n(c)) \leq s^{n-1} < \sigma$ implies that $\sigma \neq \sigma_n$. Thus, $\sigma_{n-1}(T_n(c)) \leq \sigma$ with equality when $n = 2$. This is the basis for a proof by induction on n that $\sigma_{n-1}(T_n(c)) = \sigma$. Assume that $n > 2$ and the result is true for $n - 1$. Using the interlacing property of the singular values and the induction hypothesis, we have

$$\sigma_{n-2}\big(T_n(c)\big) \geq \sigma_{n-2}\big(T_{n-1}(c)\big) = s^{n-3}\sqrt{1+c} > \sigma.$$

Therefore, only $\sigma_{n-1}(T_n(c))$ can be equal to σ.

Solution II by Leslie Foster, San Jose State University, San Jose, CA. Let $R = T_n(c)$. The squares of the singular values $\sigma_1 \geq \cdots \geq \sigma_n$ of R are the eigenvalues of $R^T R$. As in solution I, we may assume that $0 < s < 1$. The example given there shows that $\sigma = s^{n-2}\sqrt{1+c}$ is a singular value of R and $\sigma \neq \sigma_n$.

For $1 \leq k \leq n-1$, let B_k be the n by k matrix that has zero entries except for 1's on the main diagonal and -1's on the first subdiagonal. Also let $t_k = s^{2(k-1)}(1+c)$, and let $\mathcal{S}_k$ be the set of k-dimensional subspaces of Euclidean n-space $\mathbb{R}^n$. We will show that $\sigma_k \geq \sqrt{t_k}$, so that $\sigma_k > \sigma$ for $k < n-1$.

By the maxi-min characterization of singular values,

$$\sigma_k = \max_{S \in \mathcal{S}_k} \min_{y \in S} \frac{\|Ry\|}{\|y\|} \geq \min_{x \in \mathbb{R}^k} \frac{\|RB_k x\|}{\|B_k x\|}.$$

The square of the solution to this last minimization problem is the smallest eigenvalue to the generalized eigenvalue problem $Ax = \lambda Mx$, where $A = B_k^T R^T R B_k$ and $M = B_k^T B_k$. If x and λ satisfy $Ax = M\lambda x$, then it follows that $x^T(A - t_k M)x = (\lambda - t_k)x^T Mx$. Therefore, if $A - t_k M$ is symmetric semidefinite and M is symmetric positive definite, we may conclude that all the eigenvalues of $Ax = \lambda Mx$ are at least as large as t_k. Since B_k has full column rank, it follows that $M = B_k^T B_k$ is symmetric positive definite. By a straightforward calculation, $A - t_{k+1}M$ equals

$$(1+c)\begin{pmatrix} 2 & -s^2 & 0 & \cdots & 0 \\ -s^2 & 2s^2 & -s^4 & \ddots & \vdots \\ 0 & -s^4 & 2s^4 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -s^{2k} \\ 0 & \cdots & 0 & -s^{2k} & 2s^{2k} \end{pmatrix} - (1+c)s^{2k}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & 2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}.$$

From this formula, it follows that $A - t_k M$ is diagonal semidominant, is symmetric, and has nonnegative diagonal entries. It follows by known results that $A - t_k M$ is symmetric semidefinite. Thus $\sigma_k^2 \geq t_k$.

The program *Matlab* was helpful in discovering this proof. Numerical experiments helped to identify σ as an eigenvalue of $R^T R$ and to find the form of the matrix A used in the proof.

Editorial comment. The proposer notes that $T_n(c)$ is a well-known example in numerical linear algebra. The QR decomposition with column pivoting of $T_n(c)$ is itself. For $c = 0.2$ and $n = 100$, the (n, n) element of $T_n(c)$ is about .13, while its smallest singular value is about 10^{-8}. Therefore, QR decomposition with column pivoting does not reveal the near rank deficiency of this matrix. This is given as Example 5.5.1 on p. 245 of G. Golub and C. Van Loan, *Matrix Computations*, 2nd edition, The Johns Hopkins University Press, 1989. Section 8.3.1 of the same book is one reference for the well-known properties of the singular values used in the selected solutions.

Zeros of a Geometric Series with Random Signs

10351 [1993, 952]. *Proposed by Leopold Flatto and Jeffrey C. Lagarias, AT&T Bell Laboratories, Murray Hill, NJ.* Consider the random power series $f(t) = \sum_{n=0}^{\infty} \eta_n t^n$, where the η_i are drawn independently from $\{-1, 1\}$, with the probability of $\eta_i = 1$ being p for all i. (**a**) If $p = 1/2$, show that $f(t)$ has infinitely many zeros in the interval $(0, 1)$ with probability one. (**b**) What happens if $p \neq 1/2$?

Solution of (**a**) *by Richard Holzsager, The American University, Washington, DC.* We show that, with probability one, $f(t)$ oscillates between arbitrarily large positive and negative

values as t approaches 1. Since f is bounded on any interval $[0, t]$, it is enough to show that f attains arbitrarily high positive and negative values, and by symmetry, it suffices to demonstrate the positive case.

The proof depends on *Kolmogorov's zero-one law*. This deals with a situation where there is a sequence $X_1, X_2, \ldots$ of independent random variables, such as the coefficients η_n, and says that any "tail event" has probability 0 or 1. A "tail event" is one that, for any n depends only on the values of $X_n, X_{n+1}, \ldots$.

First note that, for any fixed n, changing the first n signs can change $f(t)$ by at most n. Thus, the given property is clearly a tail event.

We now show that this event does not have probability zero. Suppose M is any large integer. Then by the "gambler's ruin" result in the theory of random walks (see W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. I (third edition), Wiley, 1968, p. 347, eq. 2.8), $\sum_{i=0}^{n} \eta_i > M$ for some n with probability one. Since the polynomial $f_n(t) = \sum_{i=0}^{n} \eta_i t^i$ is continuous, it has (with probability one) a value greater than M for t close to 1. Choose such a t; by symmetry, the tail of the geometric series is nonnegative with probability at least $1/2$, so $Pr\left(f(t) \geq f_n(t)\right)$ is at least $1/2$. Thus, for $M = 1, 2, \ldots$ we have a decreasing sequence of events $\{ f(t)$ is greater than M for some $t \}$, each with probability at least $1/2$. The probability of their intersection is the limit of their probabilities, so it too is at least $1/2$.

Solution of (**b**) *by Jaime Lobo, Universidad de Costa Rica, San José, Costa Rica.* If $p \neq 1/2$, then f has only finitely many zeros in $(0, 1)$ with probability one. Indeed, f is analytic in $\mathbf{D} = \{ z : |z| < 1 \}$ and $f(0) \neq 0$, so f has only finitely many zeros in every closed interval $[0, t]$ with $0 < t < 1$. It thus suffices to show that, almost surely, $f(t)$ is of constant sign for t near 1.

Consider the function $F(z) = (1 - z)f(z)$. It too is analytic in $\mathbf{D}$. Its associated power series is $a_0 + \sum_{n>0}(a_n - a_{n-1})z^n$. Then the mth partial sum of this series, evaluated at $z = 1$, is a_m. Also, it follows from the strong law of large numbers that

$$\lim_{m\to\infty} \frac{a_0 + \cdots + a_m}{m+1} = 2p - 1$$

almost surely. In this case the Cesaro sum of the series $a_0 + \sum_{n>0}(a_n - a_{n-1})$ is also $2p - 1$ and so, from an extension of Abel's theorem, $\lim_{t\to 1} F(t) = 2p - 1$. The sign of $F(t)$, and hence that of $f(t)$ agrees with that of $2p - 1$, almost surely, for t close to 1. Indeed, $f(t)$ approaches either $+\infty$ or $-\infty$.

Editorial comment. Frank Schmidt supplied a reference to Anton Bovier and Pierre Picco, A law of the iterated logarithm for random geometric series, *Annals of Probability*, 21 (1993), 168–184, in which the asymptotic behavior of $f(t)$ as $t \to 1$ in case (**a**) is studied in fine detail.

Solved also by J. H. Lindsey II, F. Schmidt, and the proposers.

Still No Solutions

10353 [1993, 952]. *Proposed by Barry Powell, Kirkland, WA.* Show that, for any odd prime p, there do not exist nonzero integers x, y, z satisfying

$$(x, y) = 1 \qquad p \nmid xy \qquad x^6 + y^6 = z^p.$$

Composite solution by Robin J. Chapman, University of Exeter, Exeter, UK, and the proposer. If $p = 3$, the result follows from Fermat's Last Theorem for exponent 3, so assume $p \geq 5$.

As suggested in the note accompanying the statement of the problem, we use

Lemma 1. *Suppose a and b are coprime odd positive integers with $a \equiv b \pmod 4$. If m is an odd positive integer and $Q_m = Q_m(a, b) = (a^m - b^m)/(a - b)$, then Q_m is odd. If n is another odd positive integer, then the Jacobi symbols (Q_m/Q_n) and (m/n) are equal.*

Proof. See Lemma 6.1 of Chapter IV of Paolo Ribenboim, *13 Lectures on Fermat's Last Theorem*, Springer-Verlag, 1979.

Factoring the target equation yields $z^p = (x^2 + y^2)(x^4 - x^2y^2 + y^4)$. It follows that the greatest common divisor of the latter two factors divides both $3x^4$ and $3y^4$. Since $(x, y) = 1$, the greatest common divisor divides 3. But 3 cannot divide a sum of two relatively prime squares, so the factors are relatively prime. Thus $x^2 + y^2 = v^p$ and $x^4 - x^2y^2 + y^4 = w^p$ for relatively prime integers v and w. Also, x and y cannot both be odd; otherwise $x^6 + y^6$ would be congruent to 2 mod 4 and could not be a pth power. Thus $v \equiv w \equiv 1 \pmod 4$.

We now see that $3x^2y^2 = v^{2p} - w^p = (v^2 - w)Q_p(v^2, w)$. The two factors on the right are relatively prime. To see this, note that $Q_p(v^2, w) \equiv pw^p \pmod{v^2 - w}$. Then, $\gcd(w, v^2 - w) = 1$ follows from $\gcd(v, w) = 1$; and $\gcd(p, v^2 - w) = 1$ follows from $\gcd(p, xy) = 1$.

Also, since v and w are relatively prime, $v^{2p} - w^p$ is divisible by 3 only if $v^2 \equiv w \equiv 1 \pmod 3$. Thus, $Q_p(v^2, w)$ is a square and $v^2 - w$ is three times a square. Finally, since p is prime, there exists a prime q for which $(p/q) = -1$. The lemma then implies

$$-1 = (p/q) = \left(Q_p(v^2, w)/Q_q(v^2, w)\right),$$

which is impossible since $Q_p(v^2, w)$ is a square.

On the Number of Ties between Players of Equal Strength

10355 [1994, 75]. *Proposed by Joaquín Gómez Rey, I. B. "Luis Buñuel", Alcorcón (Madrid), Spain.* Two players of equal strength play a tournament consisting of $2n$ matches. Let T be the random variable that counts the number of times the score is tied during the tournament (including the initial 0-0). What is $E(T) + E(T^2)$?

Solution by Dennis P. Walsh, Middle Tennessee State University, Murfreesboro, TN. The answer is $2(n + 1)$. For $k = 0, 1, \ldots, n$, let T_{2k} be the random variable with value 1 if the score is tied after $2k$ games and value 0 otherwise. The score is tied only after an even number of games, so $T = \sum_{k=0}^{n} T_{2k}$, and

$$E(T) = \sum_{k=0}^{n} E(T_{2k}) = \sum_{k=0}^{n} \binom{2k}{k}\left(\frac{1}{4}\right)^k.$$

For $k > j$, we have

$$E(T_{2k}T_{2j}) = Pr(T_{2k} = 1 | T_{2j} = 1)\, Pr(T_{2j} = 1) = \binom{2k - 2j}{k - j}\binom{2j}{j}\left(\frac{1}{4}\right)^k.$$

Since $T_{2k}^2 = T_{2k}$, this yields

$$E(T^2) = E\left(\sum_{k=0}^{n} T_{2k}^2\right) + 2\sum_{j<k} E\left(T_{2j}T_{2k}\right) = E(T) + 2\sum_{j<k}\binom{2k - 2j}{k - j}\binom{2j}{j}\left(\frac{1}{4}\right)^k.$$

Using the well-known sum $\displaystyle\sum_{j=0}^{k}\binom{2k - 2j}{k - j}\binom{2j}{j} = 4^k$, we obtain

$$E(T)+E(T^2)=2\sum_{k=0}^{n}\binom{2k}{k}\left(\frac{1}{4}\right)^k+2\sum_{j<k}\binom{2k-2j}{k-j}\binom{2j}{j}\left(\frac{1}{4}\right)^k$$
$$=2\sum_{k=0}^{n}\left(\frac{1}{4}\right)^k\sum_{j=0}^{k}\binom{2k-2j}{k-j}\binom{2j}{j}=2(n+1).$$

Editorial comment. Numerous solvers cited W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. I (third edition), Wiley, 1968, pp. 96 and 110, where the random variable T (actually $T-1$, the number of returns to zero of the symmetric random walk) is carefully studied. The expectation and variance of this random variable are computed in J. Riordan, *Combinatorial Identities*, Wiley, 1968, p. 31-32, under the guise of the so-called Banach matchbox problem. José Luis Palacios supplied a reference to P. Kirschenhofer and H. Prodinger, "The higher moments of the number of returns of a simple random walk", *Adv. Appl. Prob.* 26 (1994), 561–563, where a generalization of this problem is studied.

Solved also by R. A. Agnew, M. H. Andreoli, D. Beckwith, P. Budney, R. J. Chapman (U. K.), D. A. Darling, R. P. Dobrow, R. Ehrenborg (Canada), R. A. Groeneveld, V. Hernández (Spain), R. Holzsager, R. D. Hurwitz, O. Krafft (Germany), J. H. Lindsey II, O. P. Lossers (The Netherlands), G. Loudner, L. E. Mattics, J. L. Palacios (Venezuela), D. E. Rauschenberg & Jian-Min Li, F. Schmidt, N. C. Singer, H. L. Stubbs, M. Vowe (Switzerland), H. Weingarten, E. A. Weinstein, A. N. 't Woord (The Netherlands), D. Zeilberger, NSA Problems Group, and the proposer.

A Sequence of Squares

10356 [1994, 75]. *Proposed by Shalosh B. Ekhad, Princeton, NJ.* Let X_n be defined by $X_0=0$, $X_1=1$, $X_2=0$, $X_3=1$, and for $n\geq 1$,

$$X_{n+3}=\frac{(n^2+n+1)(n+1)}{n}X_{n+2}+(n^2+n+1)X_{n+1}-\frac{n+1}{n}X_n.$$

Prove that X_n is the square of an integer for every $n\geq 0$.

Solution by Donald A. Darling, Newport Beach, CA. Define a sequence $\{c_n\}$ by setting $c_0=0$, $c_1=1$, and $c_{n+2}=nc_{n+1}+c_n$ for $n\geq 0$. Then $c_{n+3}=(n+1)c_{n+2}+c_{n+1}$, and $c_n=c_{n+2}-nc_{n+1}$. Squaring these two equations yields

$$c_{n+3}^2=(n+1)^2c_{n+2}^2+c_{n+1}^2+2(n+1)c_{n+2}c_{n+1}$$
$$c_n^2=c_{n+2}^2+n^2c_{n+1}^2-2nc_{n+2}c_{n+1}.$$

Eliminating $c_{n+2}c_{n+1}$ yields

$$c_{n+3}^2=\frac{(n^2+n+1)(n+1)}{n}c_{n+2}^2+(n^2+n+1)c_{n+1}^2-\frac{n+1}{n}c_n^2.$$

Since $c_2=0$ and $c_3=1$, the sequence c_n^2 satisfies the same recurrence as X_n with the same initial values, so $X_n=c_n^2$ for all $n\geq 0$.

Editorial comment. E. Sparre Anderson and Mogens Esrom Larsen noted that the sequence c_n is sequence number 704 in N. J. A. Sloane's *Handbook of Integer Sequences*. The solution of Murray S. Klamkin also begins by using this reference to identify the sequence c_n and its recurrence from $c_0, c_1, \ldots, c_6$. István Nemes noted that tools for using initial values of the sequence c_n to determine a recurrence, and then finding the recurrence satisfied by the c_n^2 are available in the *gfun* package in the Analysis part of the *Maple share library*. However, direct application of these tools shows only that c_n^2 satisfies a fourth order recurrence instead of the given third order recurrence as found in the selected solution. To complete the proof, a recurrence is derived for the difference between this sequence and the given one. It turns

out to be the fourth order recurrence constructed to give the c_n^2. Since the first four terms are seen to be zero, the whole sequence is zero.

Solved also by J. Alvarez (Spain), E. S. Andersen & M. E. Larsen (Denmark), J. Anglesio (France), B. D. Beasley, D. Beckwith, K. L. Bernstein, R. E. Bernstein, J. C. Binz (Switzerland), S. Byrd, D. Callan, R. J. Chapman (U. K.), O. Chen, J. Christopher, E. Cohen (France), J. M. Cohen, C. K. Cook, P. Cull, P. Deiermann, R. L. Doucette, J. S. Frame, Z. Franco, R. A. Groeneveld, R. Heller, R. Holzsager, M. S. Klamkin (Canada), D. R. Lepro, N. F. Lindquist, O. P. Lossers (The Netherlands), G. Loudner, J. B. Muskat & J. Schiff (Israel), I. Nemes (Austria), A. Nijenhuis, R. Richberg (Germany), R. M. Robinson, F. Schmidt & C. Forbin, S. Shaffer, M. Shemesh (Israel), N. C. Singer, M. Vowe (Switzerland), A. N. 't Woord (The Netherlands), A. Yandl, D. Zeitlin, NSA Problems Group, University of South Alabama Problem Group, and the proposer.

Another Way to be Catalan

10357 [1994, 75]. *Proposed by Ira Gessel, Brandeis University, Waltham, MA.* Define integers $a_{m,n}$ by

$$\frac{1}{1-u-v+2uv} = \sum_{m,n=0}^{\infty} a_{m,n} u^m v^n.$$

Show that $(-1)^j a_{2j,2j+2}$ is the Catalan number $\binom{2j}{j}\Big/(j+1)$.

Solution I by Chu Wenchang, Academia Sinica, Beijing, China. Consider the formal power series expansion

$$\frac{1}{1-u-v+2uv} = \frac{1}{(1-u)(1-v)}\frac{1}{1+\frac{uv}{(1-u)(1-v)}} = \sum_{k=0}^{\infty} \frac{(-1)^k u^k v^k}{(1-u)^{k+1}(1-v)^{k+1}}$$
$$= \sum_{i,j,k=0}^{\infty} (-1)^k \binom{k+i}{k}\binom{k+j}{k} u^{k+i} v^{k+j} = \sum_{m,n}^{\infty} u^m v^n \sum_{k\geq 0} (-1)^k \binom{m}{k}\binom{n}{k}.$$

We conclude that $a_{m,n} = \sum_{k\geq 0}(-1)^k\binom{m}{k}\binom{n}{k}$. This convolution is the coefficient of x^m in $(1+x)^m(1-x)^n$. With $m = 2j$ and $n = 2j+2$, this generating function becomes $(1-x^2)^{2j}(1-x)^2$, in which the coefficient of x^{2j} is

$$a_{2j,2j+2} = (-1)^j\binom{2j}{j} + (-1)^{j-1}\binom{2j}{j-1} = (-1)^j\binom{2j}{j}\frac{1}{j+1}.$$

Solution II by Bruce Sagan, Michigan State University, East Lansing, MI. Let *A-paths* be lattice paths in which each step adds $(1,0)$ or $(0,1)$ or $(1,1)$ to the current position; call these steps H, V, D, respectively. To each A-path with d diagonal steps, assign the weight $(-2)^d$. By considering the possibilities for each successive step, the coefficient of $u^i v^j$ in the generating function $\sum_{i,j\geq 0} a_{i,j} u^i v^j = (1-u-v+2uv)^{-1} = \sum_{n\geq 0}(u+v-2uv)^n$ is the sum of the weights of the A-paths from $(0,0)$ to (i,j).

Let *B-paths* be A-paths that have the same number of H-steps and V-steps and never rise above the line $y = x$. Let b_i be the sum of the weights of the B-paths from $(0,0)$ to (i,i), and let $B(z) = \sum_{i\geq 0} b_i z^i$. A B-path P of positive length is $P = (D, B\text{-path})$ or $P = (H, B\text{-path}, V, B\text{-path})$. Since the formal variable in $B(z)$ records movement in the horizontal coordinate and every D contributes a factor of -2, we obtain $B(z) = 1+(-2)zB(z)+zB(z)^2$. From the quadratic formula, $B(z) = \left(1+2z-\sqrt{1+4z^2}\right)/(2z)$.

Let $F_k(z) = \sum_{i\geq 0} a_{i,i+k} z^i$, and let *$A'$-paths* be A-paths with the same number of H-steps as V-steps. An A'-path of positive length consists of $(D, A'\text{-path})$ or $(H, B\text{-path}, V, A'\text{-path})$ or the transpose of the latter. Hence $F_0(z) = 1+(-2)zF_0(z)+2zB(z)$, and $F_0(z) = 1/(1+2z-2zB(z)) = 1\big/\left(\sqrt{1+4z^2}\right)$.

An A-path from $(0, 0)$ to $(i, i+2)$ must have the form (B-path, V, B-path, V, A'-path), where the two V's represent the first time the path touches $y = x+1$ and $y = x+2$. This yields $F_2(z) = B(z)^2 F_0(z)$.

The desired value $a_{2i,2i+2}$ is the coefficient of z^{2i} in $G(z) = (F_2(z) + F_2(-z))/2 = (-1+\sqrt{1+4z^2})/(2z^2)$. The well-known generating function for the Catalan numbers is $C(w) = \sum_{i\geq 0} C_i w^i = (1-\sqrt{1-4w})/(2w)$. Since $C(-z^2) = G(z)$, we have $a_{2j,2j+2} = (-1)^j C_j$.

Editorial comment. Several solvers used more sophisticated tools. Frank Schmidt used MacMahon's Master Theorem, Shalosh B. Ekhad used the method of WZ pairs, and Rolf Richberg used hypergeometric notation and Jacobi polynomials.

Solved also by J. Anglesio (France), K. L. Bernstein, G. A. Bookhout, A. E. Caicedo Núñez (Colombia), R. J. Chapman (U. K.), D. A. Darling, S. B. Ekhad , A. Firasath & C. C. Rousseau, R. Holzsager, D. R. Lepro, J. H. Lindsey II, O. P. Lossers (The Netherlands), G. Loudner, G. Miller (Canada), I. Nemes (Austria), R. C. Read (Canada), R. Richberg (Germany), F. Schmidt, M. Vowe (Switzerland), J. Wimp, A. N. 't Woord (The Netherlands), A. Zeleke, Howard University Combinatorics group, and the proposer.

Introducing the Eigenvalue 1

10362 [1994, 175]. *Proposed by Hans Liebeck and Anthony Osborne, University of Keele, Keele, England.* Let A be a real orthogonal matrix without eigenvalue 1. Let B be obtained from A by replacing one of its rows or one of its columns by its negative. Show that B has 1 as an eigenvalue.

Solution by Richard Holzsager, The American University, Washington, DC. If 1 is not an eigenvalue of a real orthogonal n by n matrix, then all real eigenvalues equal -1. The remaining eigenvalues occur in conjugate pairs with product 1. Therefore, if 1 is not an eigenvalue, the determinant is $(-1)^n$. Since $\det A = -\det B$, it is impossible for both matrices to have this property.

Editorial comment. Both Murray S. Klamkin and Godfrey Loudner cited the following theorem from which the solution follows immediately: If an orthogonal n-by-n matrix A has determinant 1 when n is odd or -1 when n is even, then 1 is an eigenvalue of A. (See L. Mirsky, *An Introduction to Linear Algebra*, Oxford University Press, 1972, p. 226).

F. Schmidt and Tad White each noted that this also follows from the fact that a continuous map from a sphere to itself without fixed points is homotopic to the antipodal map.

Since B is the composition of A with a reflection, this construction provides the inductive step in the proof that an n-by-n real orthogonal matrix is a product of at most n reflections. Such a result holds for more general orthogonal groups consisting of those linear transformations of an n-dimensional space over an arbitrary field of characteristic different from 2 preserving a non-degenerate quadratic form, a result known as the Cartan-Dieudonné theorem (see E. Artin, *Geometric Algebra*, Interscience, 1957, Theorem 3.20, p. 129). In addition to requiring calculations over more general fields, which appeared in some solutions of the problem 10362, the proof of the Cartan-Dieudonné theorem requires consideration of the possibility of *isotropic* vectors (vectors orthogonal to themselves). However, the result needed here is easily recovered from the statement of this more general theorem. This approach was mentioned in three solutions (by H. Guggenheimer, Allan Pedersen, and Daniel B. Shapiro).

Solved by 44 readers (including those cited) and the proposer.

More Binomial Coefficients

10363 [1994, 175]. *Proposed by Joseph M. Santmyer, California University of Pennsylvania, California, PA.* If m, n are integers satisfying $1 \le m \le n-1$, prove that

$$\binom{2n-m-1}{2n-2m-1} - \binom{n-1}{m} = \sum_k \sum_j \binom{k+j}{k} \binom{2n-m-2k-j-3}{2(n-m-k-1)}.$$

Solution by Michael Vowe, Therwil, Switzerland. We can evaluate the sum on j by equating coefficients of x^{m-1} in

$$(1-x)^{-(k+1)}(1-x)^{-(2n-2m-2k-1)} = (1-x)^{-(2n-2m-k)},$$

using the binomial theorem expansion of $(1-x)^{-a}$. We obtain for the right side of the proposed identity

$$\begin{aligned}\sum_{k=0}^{n-m-1} \binom{2n-m-k-2}{m-1} &= \sum_{k=0}^{n-m-1} \left[\binom{2n-m-k-1}{m} - \binom{2n-m-k-2}{m} \right] \\ &= \binom{2n-m-1}{m} - \binom{n-1}{m},\end{aligned}$$

which equals the left side.

Solved also by E. S. Andersen & M. E. Larsen (Denmark), J. Anglesio (France), J. C. Binz (Switzerland), G. A. Bookhout, R. J. Chapman (U. K.), H. van Haeringen (The Netherlands), R. Holzsager, O. P. Lossers (The Netherlands), G. Loudner, C. A. Minh, I. Nemes (Austria), R. C. Read (Canada), E. Schmeichel, T. White, Anchorage Math Solutions Group, NSA Problems Group, and the proposer.

A Bijection Between Sets of Permutations

10364 [1994, 176]. *Proposed by Frank Schmidt, Arlington, VA.* Let S_{2n} denote the symmetric group of degree $2n$. Let E_{2n} (respectively O_{2n}) be the set of those permutations in S_{2n} *all* of whose cycle lengths are even (respectively odd). Show that E_{2n} and O_{2n} are equinumerous by finding an explicit bijection between them.

Solution by David Callan, University of Wisconsin, Madison, WI. We say that a permutation $\pi \in S_n$ is in *standard form* if its cycles are arranged so that the smallest element in each cycle occurs in the first position and these first elements are increasing left to right (for example, $(1,5,2)(3,6,7)(4)(8)$ is in standard form). Each element of O_{2n} has an even number of cycles. Suppose $\pi \in O_{2n}$ is $\pi_1\pi_2\cdots\pi_{2k}$ in standard form. Form the corresponding $\pi' \in E_{2n}$ as follows. For $i = 1, 2, \ldots, k$, move the second element of π_{2i} to the end of π_{2i-1} unless π_{2i} is a singleton (fixed point), in which case move that element itself to the end of π_{2i-1} (and delete π_{2i}). For example, if $\pi = (1,5,2)(3,8,7)(4)(6)(9,12,10)(11)$, then $\pi' = (1,5,2,8)(3,7)(4,6)(9,12,10,11)$. Note that π' is also in standard form.

The map $\pi \to \pi'$ is the desired bijection. Its inverse is constructed as follows. Given $\pi' \in E_{2n}$ written as $\pi_1\pi_2\cdots\pi_j$ in standard form, let a denote the last element of π_1, and let b denote the first element of π_2 (if $j > 1$). If a exceeds b, then move a from π_1 so that it immediately follows b in π_2 and begin the process anew at π_3 (if present). If b exceeds a or if $j = 1$, then make a a singleton placed immediately after π_1 and begin the process anew at π_2 (if present).

Editorial comment. Letting T_{2n} be the subset of E_{2n} consisting of the permutations whose cycles all have length 2, Robert Steinberg also exhibited bijections between E_{2n} and $T_{2n} \times T_{2n}$ and between O_{2n-1} and the subset of $T_{2n} \times T_{2n}$ consisting of all (α, β) such that $\alpha(1) = \beta(1)$.

Solved also by D. Beckwith, O. P. Lossers (The Netherlands), R. Steinberg, A. N. 't Woord (The Netherlands), and the proposer.

REVIEWS

Edited by **Underwood Dudley**
Mathematics Department, De Pauw University, Greencastle, IN 46135

The Encyclopedia of Integer Sequences, by N. J. A. Sloane and Simon Plouffe. Academic Press, New York–London, 1995, xiv + 587, $44.95.

Reviewed by **Richard Guy**

John Conway calls the *Encyclopedia* "the best present I've had in years". It's a third of a century ago since Motzkin observed that the most mathematics you could get for your dollar was in Abramowitz & Stegun **[1]**, and this is probably still true today. But surely one of the best contenders for second place must be the *Encyclopedia*, the more than twice times enlarged edition of the *Handbook* **[14]**. We will refer to sequences in the *Handbook* and in the *Encyclopedia* by their **N** and **M** numbers. Not only does the *Encyclopedia* contain an enormous amount of mathematics, but it also contains what is even more important, an enormous amount of *potential* mathematics. It is hard to think of a branch of mathematics where it won't be useful, and very easy to think of other subjects where it will.

Most of us confine our use of the *Encyclopedia* to diving into the middle and looking for the sequence of our current interest, but it's well worth taking the time to read Chapter 1, which tells you how to get the best out of the book, Chapter 2 on handling a strange sequence, and Chapter 3 on further topics. For example, in Chapter 2 we find a section on transformations: exponential, logarithmic, Euler, Möbius and binomial transforms. Here are two further examples which didn't make it in time for publication: the boustrophedon transform and a transform arising from a problem of Recamán.

A fascinating Pascal-like triangle occurs in recent work of Arnol'd **[2, 3]**; start with 1; the next row starts with 0 and accumulates the members of the previous row, 0 1; we are now on the right and the next row starts there with 0 and accumulates from right to left, $1 \leftarrow 1 \leftarrow 0$; then start again with 0 and accumulate from left to right, 0 1 2 2; and so on in the manner of the plough (Fig. 1).

$$
\begin{array}{c}
1 \\
0 \rightarrow 1 \\
1 \leftarrow 1 \leftarrow 0 \\
0 \rightarrow 1 \rightarrow 2 \rightarrow 2 \\
5 \leftarrow 5 \leftarrow 4 \leftarrow 2 \leftarrow 0 \\
0 \rightarrow 5 \rightarrow 10 \rightarrow 14 \rightarrow 16 \rightarrow 16 \\
61 \leftarrow 61 \leftarrow 56 \leftarrow 46 \leftarrow 32 \leftarrow 16 \leftarrow 0 \\
0 \rightarrow 61 \rightarrow 122 \rightarrow 178 \rightarrow 224 \rightarrow 256 \rightarrow 272 \rightarrow 272
\end{array}
$$

Figure 1. Numbers of updown permutations of $12\ldots n$ ending in k.

Do you recognize the numbers in the border? The left border contains the **zig numbers**, traditionally called the **Euler numbers** or **secant numbers**; and the right border contains M2096, the **zag** or **tangent numbers**. What we have done is start with the sequence 1 0 0 0 0 ..., the coefficients in the exponential generating function for 1, and transformed it into M1492—how did Columbus get in there?— 1 1 1 2 5 16 61..., those in the exponential generating function for

$$\sec x + \tan x = 1 + \frac{1 \cdot x}{1!} + \frac{1 \cdot x^2}{2!} + \frac{2 \cdot x^3}{3!} + \frac{5 \cdot x^4}{4!} + \frac{16 \cdot x^5}{5!} + \frac{61 \cdot x^6}{6!} + \frac{272 \cdot x^7}{7!} + \frac{1385 \cdot x^8}{8!} + \frac{7936 \cdot x^9}{9!} + \cdots$$

a function (formerly?) familiar to first year calculus students on learning 'methods of integration'. More generally, Jessica Millar **[11]** discovered the simple relation between the exponential generating functions of the input and the output. For example, if we start with 1 1 1 1 1 ... (i.e., e^x) we obtain 1 1 2 4 9 24 77 ..., with e.g.f. $e^x(\sec x + \tan x)$, a near miss for M1194, the number of rhyme schemes, or M1195, the number of 2-connected planar maps. For a surprise, try feeding in the sequence 1 1 1 2 5 16 61 ... itself.

The other transformation arises from the sequence of Recamán **[12]**, namely $a_1 = 1$, and, for $n \geq 1$, $a_{n+1} = a_n/n$ or $a_n \times n$, according as n divides a_n or not. Sloane has generalized this into the transformation from $\{a_n\}$ to $\{b_n\}$ given by $b_1 = a_1$ and $b_{n+1} = \operatorname{lcm}(b_n, a_{n+1})/\gcd(b_n, a_{n+1})$. So Recamán's sequence is 1 2 6 24 120 20 140 1120 10080 1008 11088 924 12012 858 ..., while Sloane's transform of 1 2 3 4 5 6 ... is 1 2 6 6 30 5 35 280 2520 252 2772 231 3003 858 1430 ... .

One way in which the *Encyclopedia* could be made even more useful would be by including more arrays. The authors are aware of the problem. The only arrays that I found are M0663, the partitions; M1645, Pascal's (really Omar Khayyam's) triangle; M1722, which gives a way of multiplicatively encoding arrays (the rubric should read $2^1 = 2$, $2^1 3^1 = 6$, $2^1 3^2 5^1 = 90$, $2^1 3^3 5^3 7^1 = 47250$, $2^1 3^4 5^6 7^4 11^1 = 66852843750$); M3416, Euler's triangle; and M4730 & M4981, the Stirling numbers of the first & second kinds.

[As we go to press, Sloane tells me that 102 arrays have been added to the database by reading them by rows, or by diagonals if they're rectangular. In particular, Pascal's triangle comes out as

1 1 1 1 2 1 1 3 3 1 1 4 6 4 1 1 5 10 10 5 1...

and the number of partitions of n into k parts as

1 1 1 1 1 1 1 2 1 1 1 2 2 1 1 1 3 3 2 1 1 1 3 4 3 2 1 1 1 4 5 5 3 2 1 1...]

Here are a couple of examples (in addition to our Fig. 1) of what might have appeared, though three- and more-dimensional arrays present even greater problems.

Suppose that you are interested **[8, 10]** in the number of walks, $w_n(x, y)$, of n steps, each in the direction N, S, E or W, starting from the origin and ending at the lattice point (x, y), which do not stray outside the positive quadrant. The numbers of such walks form a "Pascal quarter-pyramid" (Fig. 2).

1	

1	
	1

1		
	2	
2		1

1			
	3		
5		3	
	5		1

1				
	4			
9		6		
	16		4	
10		9		1

1					
	5				
14		10			
	35		10		
35		35		5	
	35		14		1

1						
	6					
20		15				
	64		20			
84		90		15		
	140		64		6	
70		84		20		1

1							
	7						
27		21					
	105		35				
168		189		35			
	378		189		21		
294		378		105		7	
	294		168		27		1

1								
	8							
35		28						
	160		56					
300		350		70				
	840		448		56			
840		1134		350		28		
	1344		840		160		8	
588		840		300		35		1

Figure 2. The layers, $0 \le n \le 8$, of a Pascal quarter-pyramid: values of $w_n(x, y)$.

It can be shown that the entries in Fig. 2 are given by

$$w_n(x, y) = \binom{n}{r}\binom{n+2}{s} - \binom{n+2}{r+1}\binom{n}{s-1}$$

where $r = \frac{1}{2}(n + x - y)$, $s = \frac{1}{2}(n - x - y)$. Which rows, columns and diagonals are in the *Encyclopedia*? The numbers of such walks with $2k + 2$ steps from $(0,0)$ to $(0,2)$, or from $(0,0)$ to $(1,1)$, or half this last number, do not appear to be there. Of course, the number of walks from $(0,0)$ to (k, k) in $2k$ steps, is well known, and is sequence M1645.

A remarkable coincidence is that M1972, the entries $w_{2k}(0,0)$ in Fig. 2, i.e., the product of successive Catalan numbers, $c_k c_{k+1}$, is twice $w_{2k-1}(0,1)$, the number M3978, of inequivalent Hamiltonian rooted maps on $2k$ vertices; although Tutte **[15]** doesn't give the formula in that form. Is there another opportunity for a purely combinatorial proof?

For walks in the positive quadrant it's more natural and symmetrical to ask for the numbers of walks which terminate at various distances from the origin, using the "Manhattan metric", $x + y = n - 2s$. Fig. 3 shows the sums of the diagonals of Fig. 2.

w_n	n	0	1	2	3	4	5	6	7	8	9	10	11
1	0	1											
2	1		2										
6	2	2		4									
18	3		10		8								
60	4	10		34		16							
200	5		70		98		32						
700	6	70		308		258		64					
2450	7		588		1092		642		128				
8820	8	588		3024		3414		1538		256			
31752	9		5544		12276		9834		3586		512		
116424	10	5544		31680		43230		26752		8194		1024	
426888	11		56628		141570		138424		69784		18434		2048

Figure 3. Sums of diagonals of Fig. 1: values of $w'_n(x + y) = w'_n(n - 2s)$.

The entries in Fig.3 are

$$w_n'(x+y) = \binom{n+2}{s}\left[\binom{n}{s} + \binom{n}{s+1} + \cdots + \binom{n}{n-s}\right]$$
$$- \binom{n}{s-1}\left[\binom{n+2}{s+1} + \binom{n+2}{s+2} + \cdots + \binom{n+2}{n-s+1}\right]$$

Except for small values of s, the truncated binomial expansions do not seem to have a simple closed form:

$$w_n'(n) = 2^n$$
$$w_n'(n-2) = (n-2)2^n + 2$$
$$w_n'(n-4) = \tfrac{1}{2!}(n^2 - 5n + 2)2^n + n^2 + 3n - 2$$

An amusing curiosity is that $w_n'(n-2)$ is twice the genus of the $(n+2)$-dimensional cube M3874 [**13**, or see Theorem 14 in **9**], though here it may be less fruitful to look for a direct combinatorial connexion.

Of course, the first diagonal is there as M1129, and so is the third, M4723, but not the fourth. Subsequent ones are ruled out by the *Encyclopedia*'s Rule 3, since the first term which exceeds one also exceeds 999, but the online version of the *Encyclopedia* (which by now has more than 12000 entries!) now includes such sequences.

The unrelenting cascade of numbers is now relieved by numerous figures, some of which give us pictures of the actual objects being counted. Others list Hard, Disallowed and Silly sequences! Many mistakes have been corrected. N0268, in which Cayley tried to count hydrocarbons, is now supplanted by M0718. N1186 & N1635, where Cayley made some errors in connexion with his game of Mousetrap [**5**], have been corrected to M2945 & M3962, and N1423 has been extended to M3507. N1623 turned out to have some erroneous values. It should have been the same as N1622, which is now M3939; one of a wealth of examples where the same sequence occurs in quite different contexts (combination locks, evaluation of an integral, barycentric subdivisions of a simplex).

Many sequences now contain more terms than earlier, but several are crying out for similar treatment. In some cases even one more term would be a noteworthy contribution to mathematics. How many projective planes are there of order 11?

M5482: what is the number of magic squares of order 6?
M2817: what is the number of topologies on a set of 8 elements?
M1197: how many geometries (matroids) are there with 9 points?
M1585: what is the maximum kissing number [**4**] for a 10-dimensional lattice?
M3690: how many reduced latin squares are there of order 11?
M1495: how many partially ordered sets are there with 14 elements?
M0219: how many 26-dimensional unimodular lattices [**4**] are there?
M3736: what is the number of inequivalent Hadamard matrices of order 32?

Of course, we must beware of the Strong Law of Small Numbers! What is the sequence

$$2 \quad 3 \quad 5 \quad 7 \quad 11 \quad 13 \quad 17 \quad 19 \quad 23 \quad \ldots ?$$

Is it M0651 or M0652? Or even M0653? How about

$$1 \quad 3 \quad 7 \quad 19 \quad 51 \quad 141 \quad 393 \quad 1107 \quad 3139 \quad ?$$

Is the next term 8953 or 8955 [7] ? And should the sequence

$$1, 0, 1, 1, 1, 2, 2, 3, 3, 4, 5, 6, 7, 9, 10, 12, 14, 17, 19, 23, 26,$$

continue as in M0265, or as in M0266, or possibly [6] with 31, 35, 41, 46, 54, 60, 69, 78, 89, ... ?

I won't give the email address for the online version, nor that of `superseeker`, in case the network seizes up from all your enquiries. Buy the book and find out all about them.

REFERENCES

1. Milton Abramowitz and Irene A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, 1964, reprinted Dover, 1965.
2. V. I. Arno'ld, Bernoulli-Euler updown numbers associated with function singularities, their combinatorics and arithmetics, *Duke Math. J.*, **63**(1991) 537–555; *MR* **93b**:58020.
3. V. I. Arno'ld, Snake calculus and the combinatorics of the Bernoulli, Euler and Springer numbers for Coxeter groups (Russian), transl. in *Russian Math. Surveys*, **47**(1992) 1–51; *MR* **93h**:20042.
4. J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, Springer-Verlag, 1988.
5. Richard K. Guy and Richard J. Nowakowski, Mousetrap, this MONTHLY, **101**(1994) 1007–1010.
6. Richard K. Guy, The strong law of small numbers, this MONTHLY, **95**(1988), 697–712, Example 35.
7. Richard K. Guy, The second strong law of small numbers, *Math. Mag.*, **63**(1990), 3–20, Example 80.
8. Richard K. Guy, C. Krattenthaler and Bruce E. Sagan, Lattice paths, reflections and dimension-changing bijections, *Ars Combin.*, **34**(1992) 3–15.
9. Frank Harary, Topological concepts in graph theory, in Harary and Beineke, A Seminar on Graph Theory, Holt, Reinhart and Winston, New York and London, 1967, pp. 13–17.
10. Christian Krattenthaler, Counting lattice paths with a linear boundary. I. *Österreich. Akad. Wiss. Math.-Natur. Kl. Sitzungsber. II* **198**(1989) 87–107.
11. J. Millar, N. J. A. Sloane and N. E. Young, A new operation on sequences: The boustrophedon transform, *J. Combin. Theory Ser. A*, (1996) (to appear).
12. Bernardo Recamán, Problem in this MONTHLY, **102**(1995) foot of p. 924.
13. Gerhard Ringel, Über drei kombinatorische Probleme am n-dimensionalen Würfel und Würfelgitter, *Abh. Math. Sem. Univ. Hamburg*, **20**(1955) 10–19.
14. N. J. A. Sloane, *A Handbook of Integer Sequences*, Academic Press, New York, 1973.
15. W. T. Tutte, A census of Hamiltonian polygons, *Canad. J. Math.* **14**(1962) 402–417.

Department of Mathematics and Statistics
The University of Calgary
Calgary, Alberta T2N 1N4
CANADA
rkg@cpsc.ucalgary.ca

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, P. *Einstein Atomized: More Science Cartoons.* Sidney Harris. Copernicus (Imprint: Springer-Verlag), 1996, $14 (P). [ISBN 0-387-94665-9]

General, L. *The Nature of Space and Time.* Stephen Hawking, Roger Penrose. Princeton Univ Pr, 1996, ix + 141 pp, $24.95. [ISBN 0-691-03791-4] Is quantum mechanics a final theory? Can it be combined with general relativity to produce quantum gravity? In 1994, in conscious imitation of the famous Einstein-Bohr debate, Hawking and Penrose delivered a series of lectures and a capstone debate on the state of physics. Penrose, like Einstein, believes that the real world exists and that physicists exist to explain it. Hawking counters, "... I don't know what it [reality] is," and maintains that physicists exist to construct theories that accurately predict the results of measurements. SK

Mathematics Appreciation, P. *Get a Grip on Your Math.* William J. Adams. Kendall/Hunt, 1996, xiii + 256 pp, $18.95 (P). [ISBN 0-7872-1561-9] Illustrates and discusses the use and mis-use of numbers and mathematical models in everyday life. DH

Mathematics Appreciation, P. *Get a Firmer Grip on Your Math.* William J. Adams. Kendall/Hunt, 1996, vii + 290 pp, $18.95 (P). [ISBN 0-7872-1562-7] Provides further investigation and discussion questions for ideas developed in *Get a Grip on Your Math.* DH

Education. *The Role of Mathematics in Modern Engineering.* Eds: Alan K. Easton, Joseph M. Steiner. Studentlitteratur, 1996, 724 pp. [ISBN 91-44-00058-8] Proceedings of the 1st Biennial Engineering Mathematics Conference, *AEMC94*, held in July 1994 in Melbourne, Australia.

History, P, L*. *Celebrating Women in Mathematics and Science.* Ed: Miriam P. Cooney. NCTM, 1996, vii + 223 pp, $22.50 (P). [ISBN 0-87353-425-5] Intended for middle and junior high school students, these 22 biographies—from Hypatia to Dian Fossey to Mary Ellen Rudin—are captivating readings for everyone. Each biography ends with suggested readings. DH

History, S, P*, L*. *Sources of Hyperbolic Geometry.* John Stillwell. History of Math., V. 10. AMS and London Math Society, 1996, ix + 153 pp, $39. [ISBN 0-8218-0529-0] Introductory commentaries (with lists of references) and English translations of papers on hyperbolic geometry by Beltrami, Klein, and Poincaré. A great resource for those interested in the development of non-Euclidean geometry. JNC

History, L. *Modern Algebra and the Rise of Mathematical Structures.* Leo Corry. Science Networks: Historical Studies, V. 17. Birkhäuser Boston, 1996, 460 pp, $139. [ISBN 0-8176-5311-20 The development of the idea of mathematics as a study of structures as exemplified by ideal theory and category theory. Concentrates on contributions of Dedekind, Hilbert, Fraenkel, Noether, Ore, and the Bourbaki group. DB

Logic, P. *Logic and Algebra.* Eds: Aldo Ursini, Paolo Agliano̊. Lect. Notes in Pure & Appl. Math., V. 180. Marcel Dekker, 1996, xv + 702 pp, $175 (P). [ISBN 0-8247-9606-

3] Proceedings of a 1994 conference in Siena (Italy) honoring Roberto Magari.

Combinatorics, P. *Computational and Constructive Design Theory.* Ed: W.D. Wallis. Math. & Its Applic., V. 368. Kluwer Academic, 1996, xiv + 357 pp, $165. [ISBN 0-7923-4015-9] 11 papers (2 tutorial) on computational techniques in constructive design theory.

Number Theory, T(18: 1). *Additive Number Theory: The Classical Bases.* Melvyn B. Nathanson. Grad. Texts in Math., V. 164. Springer-Verlag, 1996, xiv + 342 pp, $49.95. [ISBN 0-387-94656-X] Thorough account of Waring's problem (number of representations of an integer as a sum of a specified number of specified powers) and Goldbach's conjecture (the existence of a representation of an even integer as a sum of two primes). Describes the circle method and sieve techniques. With historical comments and exercises. DB

Number Theory, T(18: 1). *Additive Number Theory: Inverse Problems and the Geometry of Sumsets.* Melvyn B. Nathanson. Grad. Texts in Math., V. 165. Springer-Verlag, 1996, xiv + 293 pp, $49.95. [ISBN 0-387-94655-1] A variety of results in additive number theory including inverse problems (determine the original set from the sums of elements) and problems involving congruence classes, lattice points, graphs, and combinatorics. With historical notes and exercises. DB

Linear Algebra, T(16–18: 1, 2), S, P, L. *A Polynomial Approach to Linear Algebra.* Paul A. Fuhrmann. Universitext. Springer-Verlag, 1996, xiii + 360 pp, $39 (P). [ISBN 0-387-94643-8] Not an introductory linear algebra text; accessible only after previous courses in linear algebra and algebraic structures. Supplies an interesting gateway to several non-standard topics including shift operators, quadratic forms, system theory, and Hankel norm approximation. Formal and rigorous; linear transformations are studied by looking at module structure induced by rings of polynomials. Exercises are largely theoretical, little routine computation. JS

Algebra, T(18), P. *Foundations of Quantum Group Theory.* Shahn Majid. Cambridge Univ Pr, 1995, xix + 607 pp, $100. [ISBN 0-521-46032-8] Emphasizes the algebraic theory of quantum groups and (more generally) Hopf algebras. Provides detailed proofs, thorough motivations, and a diverse choice of topics. TH

Algebra, T(18), P. *Cohomology of Drinfeld Modular Varieties, Part I: Geometry, Counting of Points and Local Harmonic Analysis.* Gérard Laumon. Stud. in Adv. Math., V. 41. Cambridge Univ Pr, 1996, xiii + 344 pp, $64.95. [ISBN 0-521-47060-9] Self-contained development of the function field analogs of Shimura varieties over number fields. TH

Algebra, P. *The Group Fixed by a Family of Injective Endomorphisms of a Free Group.* Warren Dicks, Enric Ventura. Contemp. Math., V. 295. AMS, 1996, ix + 81 pp, $19 (P). [ISBN 0-8218-0564-9] An algebraic proof of the Bestvina-Handel Theorem, which gives an upper bound on the rank of the fixed group of an automorphism of a free group. TH

Algebra, P. *Nilpotent Lie Algebras.* Michel Goze, Yusupdjan Khakimdjanov. Math. & Its Applic., V. 361. Kluwer Academic, 1996, xv + 336 pp, $170. [ISBN 0-7923-3932-0]

Algebra, P. *Rings, Groups, and Algebras.* Eds: X.H. Cao, *et al.* Lect. Notes in Pure & Appl. Math., V. 181. Marcel Dekker, 1996, viii + 332 pp, $150 (P). [ISBN 0-8247-9733-7] Survey articles and recent research results. Summarizes the most significant developments in rings and algebras made in China since the 1950's.

Calculus, S(13). *Calculus: A Lab Course with MicroCalc.* Harley Flanders. Springer-Verlag, 1996, xi + 332 pp, $39.95 (P). [ISBN 0-387-94496-6] A lean calculus text using the author's software. PF

Calculus, S?(13). *A Concise Introduction to Calculus.* W.Y. Hsiang. Ser. on Univ. Math., V. 6. World Scientific, 1995, vii + 157 pp, $21 (P); $39. [ISBN 981-02-1901-6; 981-02-1900-8] A meager attempt to explain calculus in a minimum number of pages. PF

Calculus, T(13: 1). *Brief Calculus with Applications in Business and the Social and Life Sciences.* Daniel C. Alexander. H & H Pub, 1996, x + 806 pp, $39.95. [ISBN 0-943202-51-5] Standard textbook for business and liberal arts students. Makes little attempt to do anything modern. PF

Calculus, S*(13). *Maple V Calculus Labs, Second Edition.* Abi Fattahi. Brooks/Cole, 1996, ix + 109 pp, $16.95 (P). [ISBN 0-534-26208-2] An excellent supplement. The labs are clearer than in the previous edition and include new examples. PF

Real Analysis, T*(15–17: 1, 2), S*, L. *Advanced Calculus: A Course in Mathematical Analysis.* Patrick M. Fitzpatrick. PWS, 1996, xix + 555 pp, $86.25. [ISBN 0-534-92612-6] A clear, readable, comprehensive introduction to a wide variety of topics in real analysis, rich with examples and exercises. Culminates with

integral formulas of Green and Stokes in the plane and space. PZ

Complex Analysis, T(18: 1). *Complex Analysis and Special Topics in Harmonic Analysis.* Carlos A. Berenstein, Roger Gay. Springer-Verlag, 1995, x + 482 pp, $79. [ISBN 0-387-94411-7] Boundary values of holomorphic functions, ideal theory in algebras of entire functions, the G-transform, summation methods, and questions in harmonic analysis such as how to recover a continuous function from its averages over intervals of different lengths. DB

Partial Differential Equations, T(18: 2), L. *Partial Differential Equations.* Michael E. Taylor. Springer-Verlag, 1996, $69 each. *I: Basic Theory,* Appl. Math. Sci., V. 115, xxi + 563 pp, [ISBN 0-387-94653-5]; *II: Qualitative Studies of Linear Equations,* Appl. Math. Sci., V. 116, xxi + 528 pp, [ISBN 0-387-94651-9]; *III: Nonlinear Equations,* Appl. Math. Sci., V. 117, xxi + 610 pp. [ISBN 0-387-94652-7] (*I: Basic Theory* is available in paperback as V. 23 in the Texts in Applied Mathematics series.) Good introduction to PDE's. Thorough classical coverage, though some important modern topics (e.g., solitons) are missing. Many exercises. JO

Dynamical Systems, P. *Dynamical Systems of Algebraic Origin.* Klaus Schmidt. Progress in Math., V. 128. Birkhäuser Boston, 1995, xviii + 310 pp, $89. [ISBN 0-8176-5174-8] The systems studied are homomorphisms of $\mathbb{Z}^d$ to the automorphisms of compact (usually abelian) groups. SK

Dynamical Systems, P. *Algorithms, Fractals, and Dynamics.* Ed: Y. Takahashi. Plenum Pr, 1995, viii + 227 pp, $85. [ISBN 0-306-45127-1] Papers from symposia held in 1995 at the Fujisaki Institute of Hayashibara Biochemical Laboratories, Inc., and at Kyoto University.

Dynamical Systems, P. *Dynamics in Several Complex Variables.* John Erik Fornæss. CBMS Reg. Conf. Ser. in Math., No. 87. AMS, 1996, vii + 59 pp, $19 (P). [ISBN 0-8218-0317-4] Ten brief, clear, readable lectures (from a 1994 CBMS conference held in Albany) comprise an introduction to recent literature and open problems in higher-dimensional complex dynamics. Lectures offer motivation rather than complicated proofs—the author believes that "mathematicians can read arbitrarily complicated material [for themselves] once they are motivated." PZ

Dynamical Systems, T(14–15: 1), C, L. *A First Course in Discrete Dynamical Systems, Second Edition.* Richard A. Holmgren. Universitext. Springer-Verlag, 1996, xv + 223 pp, $29.95 (P). [ISBN 0-387-94780-9] Technical, yet accessible introduction to iterated functions; requires only calculus. Reorganization of *First Edition* (TR, January 1995) makes metric spaces and symbolic dynamics optional. RM

Operator Theory, P. *Integral Equations with Difference Kernels on Finite Intervals.* Lev A. Sakhnovich. Operator Theory: Adv. & Applic., V. 84. Birkhäuser Boston, 1996, vi + 177 pp, $94. [ISBN 0-8176-5267-1]

Operator Theory, P. *Toeplitz Operators and Index Theory in Several Complex Variables.* Harald Upmeier. Operator Theory: Adv. & Applic., V. 81. Birkhäuser Boston, 1996, 481 pp, $161. [ISBN 0-8176-5282-5]

Operator Theory, P. *Functional Calculus of Pseudodifferential Boundary Problems, Second Edition.* Gerd Grubb. Prog. in Math., V. 65. Birkhäuser Boston, 1996, viii + 522 pp, $89.50. [ISBN 0-8176-3738-9] *First Edition,* TR, November 1987.

Analysis, T*(14: 2), L*. *Mathematical Analysis: An Introduction.* Andrew Browder. Undergrad. Texts in Math. Springer-Verlag, 1996, xiv + 333 pp, $39. [ISBN 0-387-94614-4] First third is a careful development of standard topics from calculus including series, continuity, and Riemann integration. Middle third is topology with applications (e.g., geodesics in compact metric spaces). Final third is about calculus on manifolds, including differential forms, and Brouwer's fixed point theorem. An impressive list of topics in a small, reasonably priced book. Many exercises and a useful bibliography. TAV

Analysis, P. *Lecture Notes in Control and Information Sciences–219: ICASO '96: Images, Wavelets and PDEs.* Eds: Marie-Odile Berger, *et al.* Springer-Verlag, 1996, xv + 359 pp, $63 (P). [ISBN 3-540-76076-8] Proceedings of the 12th International Conference on Analysis and Optimization of Systems (June 1996) in Paris, France.

Algebraic Geometry, P. *Algorithms in Algebraic Geometry and Applications.* Eds: Laureano González-Vega, Tomás Recio. Progress in Math., V. 143. Birkhäuser Boston, 1996, ix + 399 pp, $98.50. [ISBN 0-8176-5274-4] 20 papers from the **MEGA-94** Conference at the University of Cantabria (Spain).

Algebraic Geometry, P. *Moduli of Vector Bundles.* Ed: Masaki Maruyama. Lect. Notes in Pure & Appl. Math., V. 179. Marcel Dekker, 1996, viii + 305 pp, $135 (P). [ISBN 0-8247-9738-8] 20 papers from the 35th Taniguchi International Symposium (1994).

Differential Geometry, P. *Fundamental*

Groups of Compact Kähler Manifolds. J. Amorós, *et al.* Math. Surveys & Mono., V. 44. AMS, 1996, xi + 140 pp. [ISBN 0-8218-0498-7]

Differential Geometry, P. *Manifolds and Geometry.* Eds: Paolo de Bartolomeis, Franco Tricerri, Edoardo Vesentini. Cambridge Univ Pr, 1996, ix + 321 pp, $59.95. [ISBN 0-521-56216-3] Proceedings of a 1993 conference held in Pisa to honor Eugenio Calabi.

Differential Geometry, T(17–18: 1), P. *Riemannian Geometry and Geometric Analysis.* Jürgen Jost. Springer-Verlag, 1995, xi + 401 pp, $54 (P). [ISBN 0-387-57113-2] Introductory text on geometric and analytic methods in study of Riemannian manifolds. Reasonably self-contained. Nonlinear analysis techniques are introduced early and used throughout. Uses both invariant global and tensor notation. RM

Differential Geometry, T(17–18: 1), S, P. *An Introduction to Lorentz Surfaces.* Tilla Weinstein. Expos. in Math., V. 22. Walter de Gruyter, 1996, xiii + 213 pp, DM 168. [ISBN 3-11-014333-X] Lorentz surfaces (manifolds provided with set of metrics conformally equivalent to an indefinite Lorentz metric) are more subtle and complicated than their Riemann surface analogs, and have emerged as a tool in relativity theory. Introduction to current research in 2-dimensional Lorentz geometry, comparison with Euclidean and Minkowski 3-spaces. RM

Geometry, T(16), P, L. *Minkowski Geometry.* A.C. Thompson. Ency. of Math. & Its Applic., V. 63. Cambridge Univ Pr, 1996, xvi + 346 pp, $59.95. [ISBN 0-521-40472-X] Coverage of topological properties of Minkowski spaces (i.e., finite dimensional normed linear spaces), characterizations of Euclidean space among normed spaces, area and volume in normed spaces, and trigonometry in Minkowski spaces. Chapters end with historical notes. Lists 50 unsolved problems. JNC

Topology, P. *Compact Projective Planes With an Introduction to Octonion Geometry.* Helmut Salzmann, *et al.* Expos. in Math., V. 21. Walter de Gruyter, 1995, xiii + 688 pp, DM 258. [ISBN 3-11-011480-1] 40 years of research on topological projective planes, collineation groups, and incidence geometry. Approach connects group theory to geometry. JD

Topology, P. *Categorical Topology.* Ed: Eraldo Giuli. Kluwer Academic, 1996, v + 280 pp, $130. [ISBN 0-7923-4049-3] Proceedings of a 1994 workshop at the University of L'Aquila. Some of the papers are reprinted from *Applied Categorical Structures* **4** (1996).

Operations Research, P, L. *Queueing Systems: Problems and Solutions.* Leonard Kleinrock, Richard Gail. Wiley, 1996, ix + 227 pp, $34.95 (P). [ISBN 0-471-55568-1] All the problems from *Queueing Systems, Volume 1: Theory* by L. Kleinrock (Wiley-Interscience, 1976) and their solutions. A brief, introductory section is "A Queueing Theory Primer."

Optimal Control, P. *Mathematical Theory of Control Systems Design.* V.N. Afanas'ev, V.B. Kolmanovskii, V.R. Nosov. Math. & Its Applic., V. 341. Kluwer Academic, 1996, xxiii + 668 pp, $279. [ISBN 0-7923-3724-7]

Probability, T(15–16: 1). *Probability and Information: An Integrated Approach.* David Applebaum. Cambridge Univ Pr, 1996, xiii + 212 pp, $24.95 (P); $69.95. [ISBN 0-521-55528-0; 0-521-55507-8] Concise presentation of core material on both probability and information; defines probability as a measure on a Boolean algebra. Suggestions for further reading conclude each chapter. RSK

Elementary Statistics, T(13–14: 1, 2). *General Statistics, Third Edition.* Warren Chase, Fred Bown. Wiley, 1997, xv + 713 pp, $67.95. [ISBN 0-471-05584-0] Changes (*Second Edition*, TR, February 1993) include more use of exploratory data analysis, an earlier introduction to descriptive aspects of regression and correlation, a re-worked probability chapter; variance tests now relegated to an appendix. Includes data from the Framingham Heart Study, and problem sets based on it. Examples and exercises include output from a variety of statistical packages. RSK

Mathematical Statistics, T(18: 1, 2), P. *Multivariate Statistical Analysis.* Narayan C. Giri. Stat.: Textbooks & Mono., V. 149. Marcel Dekker, 1996, xii + 378 pp, $135. [ISBN 0-8247-9338-2] Culmination of the author's extensive research and teaching experience. Emphasizes the invariance approach. Primarily theoretical; few examples and exercises based on real data. RSK

Mathematical Statistics, T(16–17: 2, 3). *Probability and Statistical Inference.* Robert Bartoszyński, Magdalena Niewiadomska-Bugaj. Ser. in Prob. & Stat. Wiley, 1996, xvi + 826 pp, $59.95. [ISBN 0-471-31073-5] Solid text stressing comprehension over skill acquisition. Contains no computer references and does not talk down to students. Includes much extra material for mathematically-minded students, such as Lebesgue and Stieltjes integrals and chapters on Markov chains and discrimination. RSK

Statistical Methods, P. *MSI–2000: Multivariate Statistical Analysis in Honor of Profes-*

sor Minoru Siotani. Eds: Takesi Hayakawa, Makoto Aoshima, Kunio Shimizu. Amer. Journ. of Math. & Management Sci., V. 15, Nos. 3 & 4. 1995, 237 pp, $125 (P). [ISBN 0-935950-38-9] Proceedings of a 1995 conference at the University of Hawaii.

Programming, P. *Java in a Nutshell: A Desktop Quick Reference for Java Programmers.* David Flanagan. O'Reilly & Associates, 1996, xix + 438 pp, $14.95 (P). [ISBN 1-56592-183-6]

Computer Systems, P. *CGI Programming on the World Wide Web.* Shishir Gundavaram. O'Reilly & Associates, 1996, xiv + 433 pp, $29.95 (P). [ISBN 1-56592-168-2]

Computer Systems, P, L. *HTML: The Definitive Guide.* Chuck Musciano, Bill Kennedy. O'Reilly & Associates, 1996, xx + 385 pp, $27.95 (P). [ISBN 1-56592-175-5]

Computer Science, T(15: 1). *An Introduction to High-Performance Scientific Computing.* Lloyd D. Fosdick, *et al.* Sci. & Eng. Comput. MIT Pr, 1996, xxiii + 760 pp, $55. [ISBN 0-262-06181-3] Concrete introduction to tools (including UNIX, Fortran, MATLAB, etc.), algorithms, and applications of high-performance computing. Applications include molecular dynamics, advection, and tomography. JO

Computer Science, T(15–16: 1), S, L. *Algebraic Semantics of Imperative Programs.* Joseph A. Goguen, Grant Malcolm. Found. of Comp. Ser. MIT Pr, 1996, ix + 228 pp, $32. [ISBN 0-262-07172-X] Nice introduction to science of reasoning about programs. Goals: enable undergraduates to better understand the semantics of programs; develop intuitions about programming; rigorously verify properties of programs. Algebraic variation of the axiomatic approach; uses equational logic and OBJ (functional metalanguage, with equations as statements, proofs as computations) for expressing and proving program properties. RM

Applications (Engineering), T(17: 2). *Mathematical Analysis in Engineering: How to Use the Basic Tools.* Chiang C. Mei. Cambridge Univ Pr, 1995, xvii + 461 pp. [ISBN 0-521-46053-0] Includes all standard introductory topics (e.g., Fourier series, Green's functions, transforms) as well as chapters on perturbation methods and symbolic computation. Topics are introduced and explained through physical problems and examples. SK

Applications (Fluid Mechanics), P. *Advances in Multi-Fluid Flows.* Eds: Yuriko Y. Renardy, *et al.* SIAM, 1996, xvi + 432 pp, $75 (P). [ISBN 0-89871-377-3] Proceedings of the 1995 Conference on Multi-Fluid Flows and Interfacial Instabilities at the University of Washington.

Applications (Mechanics), P. *Hamiltonian Dynamics and Celestial Mechanics.* Eds: Donald G. Saari, Zhihong Xia. Contemp. Math., V. 198. AMS, 1996, vii + 240 pp, $72 (P). [ISBN 0-8218-0566-5] Proceedings of a 1995 AMS-IMS-SIAM Joint Summer Research Conference at the University of Washington.

Applications (Systems Theory), P. *Control of Uncertain Sampled-Data Systems.* Geir E. Dullerud. Systems & Control: Found. & Applic. Birkhäuser Boston, 1996, xiv + 177 pp, $42.50 (P). [ISBN 0-8176-3851-2]

Applications (Systems Theory), P. *Modelling and Optimization of Distributed Parameter Systems Applications to Engineering.* Eds: Kazimierz Malanowski, Zbigniew Nahorski, Małgorzata Peszyńska. Chapman & Hall, 1996, x + 387 pp, $100. [ISBN 0-412-72700-5] Papers from a 1995 conference held in Poland.

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–214: Recent Advances in Control and Optimization of Manufacturing Systems.* Eds: George Yin, Qing Zhang. Springer-Verlag, 1996, 222 pp, $54 (P). [ISBN 3-540-76055-5] 8 papers on optimal production planning, scheduling and improvability, and approximate optimality and robustness.

Applications, S, P*. *A Probabilistic Analysis of the Sacco and Vanzetti Evidence.* Joseph B. Kadane, David A. Schum. Ser. in Prob. & Stat. Wiley, 1996, xvi + 366 pp, $49.95. [ISBN 0-471-14182-8] Uses an elaborate case study (the celebrated murder trial and appeals of anarchists Sacco and Vanzetti, convicted and executed) to show "how modern probabilistic methods can be employed in the study of complex inferences based on masses of evidence." Includes synopses of 28 Wigmore trial and post-trial evidence charts given in an appendix, and probabilistic analyses of the evidence. Thorough discussion of all aspects of the methodology. Conclusion: Vanzetti innocent, Sacco's guilt not proven. RSK

Reviewers

DB: David Bressoud, Macalester; JNC: Judith N. Cederberg, St. Olaf; PF: Paul Froeschl, Macalester; TH: Tom Halverson, Macalester; DH: Deanna Haunsperger, Carleton; SK: Steve Kennedy, Carleton; RSK: Richard S. Kleber, St. Olaf; RM: Richard Molnar, Macalester; JO: Jeff Ondich, Carleton; JS: John Schue, Macalester; TAV: Theodore A. Vessey, St. Olaf; PZ: Paul Zorn, St. Olaf.

THE AUTHORS

BRYAN SHADER received his Ph.D. from the University of Wisconsin–Madison in 1990. He wrote his dissertation on tournament matrices under the direction of Richard Brualdi. He was a post-doctoral fellow at the Institute for Mathematics and its Applications during their year on Applied Linear Algebra in 1991. He is an Associate Professor of Mathematics at the University of Wyoming. His primary research interests involve combinatorics and linear algebra.

CHANYOUNG LEE SHADER received her Ph.D. from the University of Wisconsin–Madison in 1992. Her dissertation was supervised by Georgia Benkart and concerned the representation theory of Lie algebras. She is an Assistant Professor of Mathematics at the University of Wyoming. Her current research interests lie in the use of the combinatorics of partitions and tableaux in representation theory of Lie superalgebras.

RALPH H. BUCHHOLZ spent most of his youth in Teralba with numerous sojourns to the surrounding beaches to bodysurf. In 1981 he received an Honours degree in Mathematics from Newcastle University. Ralph then became the Distribution Analyst of the Natural Gas Company. Four years later the lure of research drew him back to Newcastle University and in 1989 he completed a Ph.D. in Mathematics. Since 1991 he has worked for the Crypto-Mathematics Research Group while maintaining an avid interest in the theory of elliptic curves, the music of Kate Bush, and the Stone of Rosetta.

RANDALL L. RATHBUN has been fascinated by mathematics since first reading in the seventh grade how the Lehmers factored large numbers. He credits Martin Gardner for steering his interests towards the rational cuboid problem and especially thanks his coauthor and Richard K. Guy for persuading him to obtain his bachelors degree in mathematical sciences at California State University San Marcos in 1994. He has taught science lessons for the K.I.D.S. program at an elementary school and algebra for the NSF Math Renaissance at a middle school. He hopes to awaken a triangle mini-renaissance in the math community and obtain a masters degree shortly.

TAO KAI LAM received his Ph.D. from MIT under Richard P. Stanley in 1995. He did his undergraduate in the National University of Singapore, where he is now a lecturer. His area of research is combinatorics. Born in Singapore in 1966, he found out that he had spent 21 years of his life studying in one school or another. This makes the transition from one end of the lecture hall to the other a big challenge. He now indulges in chewing gum whenever he is out of the country.

RONALD PRATHER completed his Ph.D. in 1969 at Syracuse University. He has been a member of the faculty at the University of Vermont, The University of Denver, and Syracuse University. He holds the Caruth Distinguished Professorship in Computer Science at Trinity University in San Antonio, TX. His research interests include programming languages and software metric theory. He is also the moderator of the Alec Wilder Mailing List on the Internet, a group devoted to discussions of the life and work of this uncommon American composer.

R. BRUCE RICHTER is interested in graph theory, combinatorics, and the topology of surfaces. He received his Ph.D. in 1983 from the University of Waterloo, Canada and has taught at Utah State University, The Ohio State University, and The U.S. Naval Academy. He has been at Carleton since 1988.

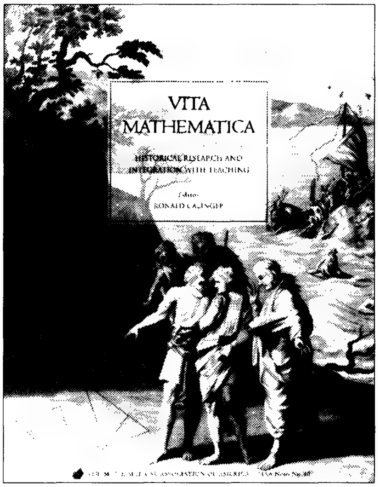
CARSTEN THOMASSEN received his Master's degree in 1972 at University of Aarhus, Denmark, and his Ph.D. in 1976 at University of Waterloo, Canada. Since 1981 he has been a professor of mathematics at the Technical University of Denmark. He is coeditor-in-chief of the *Journal of Graph Theory* and he is a member of the Royal Danish Academy of Sciences and Letters.

JOHN HOLTE liked to add long columns of numbers when he was in first grade. Subsequently he earned his B.S. in mathematics at Caltech and his Ph.D. at the University of Wisconsin. He has taught at Augustana College–Rock Island, Rensselaer Polytechnic Institute, and Gustavus Adolphus College. His research interests lie in probability theory, but after teaching courses in applied combinatorics and fractal geometry and chairing the 1990 Nobel Conference on chaos, he began investigating a multifractal description of multinomial coefficient divisibility and stumbled upon the "amazing matrix."

WALTER NEF was born in 1919 in Winterthur. He received his Ph.D. in Mathematics from the University of Zurich in 1942. From 1944 to 1948 he taught at the University of Fribourg, with a break in 1946 for a stay at Brown University. In 1948 he was offered a chair at the University of Bern. In 1956/7 he visited the National Bureau of Standards in D.C., where he got acquainted with applied mathematics and computer sciences. In Bern he founded the Institute for Applied Mathematics, which he headed until 1979. He became emeritus in 1984. His fields of interest are combinatorial and computational geometry.

RICHARD GUY's only formal mathematical education was at Warwick School and Cambridge University. He has taught at all levels from kindergarten to postgraduate, and in many parts of the world: Britain, Singapore, India, Canada. Embarrassed by having a senior professor without a doctorate, The University of Calgary gave him an honorary degree in 1991. He still visits the mountains, both summer and winter.

THE MATHEMATICAL ASSOCIATION OF AMERICA



Vita Mathematica

Historical Research and Integration with Teaching

Ronald Calinger, Editor

The use of the history of mathematics in the teaching of mathematics at all levels is an idea whose time has come. To use history in the teaching of undergraduate mathematics, the instructor must be familiar with the history as well as the mathematics. *Vita Mathematica* will enable college teachers to learn the relevant history of various topics in the undergraduate curriculum and help them incorporate this history in their teaching.

For example, should calculus be approached from a geometric or an algebraic point of view? The book shows us how two important eighteenth century mathematicians, Colin Maclaurin and Joseph-Louis Lagrange, understood the calculus from these different standpoints and how their legacy is still important in teaching calculus today. We also learn why Lagrange's algebraic approach dominated teaching in Germany in the nineteenth century. Some of the reasons for this are related to the appropriate foundations of the calculus, and so the book traces the ancient history of one of the possible foundations, the concept of indivisibles. Even though we generally do not use this concept formally today, many ideas for a heuristic approach to the calculus can be developed out of his study.

Vita Mathematica contains numerous other articles dealing with calculus, with algebra, combinatorics, graph theory, and geometry, as well as more general articles on teaching courses for prospective teachers. This volume, then, demonstrates that the history of mathematics is no longer tangential to the mathematics curriculum, but in fact deserves a central role.

Catalog Code: NTE40/JR
350 pp., Paperbound, 1996, ISBN 0-88385-097-4
List: $34.95 MAA Member: $29.00

Phone in Your Order Now! ☎ 1-800-331-1622

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders:** $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ______________________

Address ______________________

City ____________ State _____ Zip ________

Phone ______________________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	NTE40/JR	______	______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling ______

TOTAL ______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

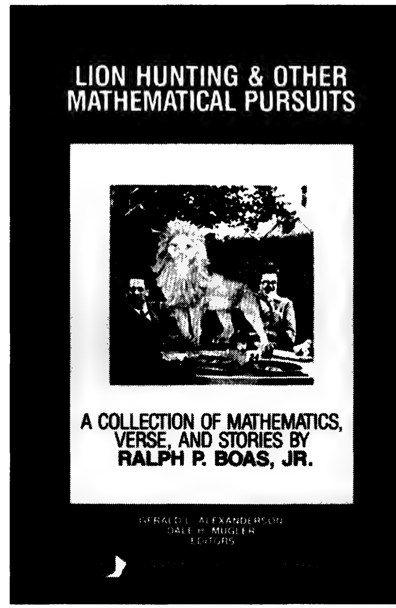
Signature ______________________

Lion Hunting

and Other Mathematical Pursuits

A Collection of Mathematics, Verse, and Stories
by Ralph P. Boas, Jr.

Gerald L. Alexanderson and
Dale H. Mugler, Editors



I highly recommend Lion Hunting and Other Mathematical Pursuits to high school mathematics clubs, mathematics teachers of all levels, and anyone interested in mathematics. Perhaps the most important features of this book is how it subtly makes the reader aware of the nature of mathematics.

– The Mathematics Teacher

As a young man at the Institute for Advanced Study in Princeton, Ralph Philip Boas, Jr., together with a group of other mathematicians, published a light-hearted article on the "mathematics of lion hunting" under a pseudonym (1938). This sparked a sequence of articles on the topic, several of which are drawn together in this book.

Lion Hunting includes an assortment of articles that show the many facets of this remarkable mathematician, editor, writer, and teacher. Along with a variety of his lighter mathematical papers, the collection includes Boas' verse and short stories, many of which are appearing for the first time. Anecdotes and recollections of his numerous experiences and of his work and meetings with many distinguished mathematicians and scientists of his day are also included as well as photographs taken by Boas of Hardy, Littlewood, Besicovitch, Weil, and others.

The mathematical articles in this collection cover a range of topics. They include articles on infinite series, the mean value theorem, indeterminate forms, complex variables, inverse functions, extremal problems for polynomials and more. A special section of this book is devoted to articles about the teaching of mathematics, with titles such as "Calculus as an experimental science" and "Can we make mathematics intelligible?"

Boas's wit and playful humor are reflected in the verses included in this collection. The verses reflect the phases of his career as author, editor, teacher, department chair, and lover of literature. A section of the book describes the feud that Boas supposedly had with Bourbaki. Also included are many amusing anecdotes about famous mathematicians.

320 pp., Paperbound, 1995, ISBN 0-88385-323-X
List: $39.95 MAA Member: $28.95
Catalog Code: DOL-15

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
P.O. Box 91112, Washington, DC 20090-1112
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
____	DOL-15	____	____
____	____	____	____
		TOTAL	____

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires __/__

Signature ______________________

THE MATHEMATICAL ASSOCIATION OF AMERICA

Julia

a life in mathematics

Constance Reid

Constance Reid, an established writer about mathematicians, has written an excellent and loving book, about her sister Julia Robinson, the mathematician. The author has written that she wants the book to be one for all age groups and she has succeeded admirably in making it so...Julia wanted to be known as a mathematician, not a woman mathematician and rightly so! However, she was, and is, a wonderful role model for women aspiring to be mathematician. What a great gift this book would be!
—Alice Schafer, Former President, AWM

This book is a small treasure, one which I want to share with all my mathematical friends. The assembly of several articles and additional photos and remarks provides the image of a mathematician of extraordinary taste, tenacity and generosity.... Julia Robinson broke ground in displaying the deep connections between number theory and logic. Her results have led to a very active area today, making the appearance of this book very timely. Her work and her example are however timeless and I can think of no better advice to give a young mathematician, either in how to do mathematics. or how to behave in mathematics, than: "Be like Julia!"
—Carol Wood, Deputy Director, MSRI

In high school Julia Bowman stood alone as the only girl—and the best student—in her junior and senior math classes. She had only one close friend and no boyfriends. Although she was to learn (from E. T. Bell's *Men of Mathematics*) that there are such people as mathematicians, her ambition was merely to get a job teaching mathematics in high school.

At great sacrifice her widowed stepmother sent her to the University of California at Berkeley to obtain the necessary teaching credentials. But at Berkeley, in a society of mathematicians, she discovered herself. She was not the duckling that didn't belong, but a swan. There was also a prince at Berkeley, a brilliant young assistant professor named Raphael Robinson. Theirs was to be a marriage that would endure until her death in 1985.

Julia is the story of the life of Julia Bowman Robinson, the gifted and highly original mathematician who during her lifetime was recognized in ways that no other woman mathematician had been recognized up to that time. In 1976 she became the first woman mathematician elected to the National Academy of Sciences and in 1983 the first woman elected president of the American Mathematical Society.

This unusual book, profusely illustrated with previously unpublished personal and mathematical memorabilia, brings together in one volume the prizewinning "Autobiography of Julia Robinson" by her sister, the popular mathematical biographer Constance Reid, and three very personal articles about her work by outstanding mathematical colleagues.

All royalties from sales of this book will go to fund a Julia Robinson Prize in Mathematics at the high school from which she graduated.

Catalog Code: JULIA/JR
136 pp., Hardbound, 1996, ISBN 0-88385-520-8
List: $27.00 MAA Member: $20.00

Phone in Your Order Now!* ☎ *1-800-331-1622

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows **USA orders (shipped via UPS):** $2 95 for the first book, and $1 00 for each additional book **Canadian orders:** $4 50 for the first book and $1 50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route We do not ship via UPS into Canada unless the customer specially requests this service Canadian customers who request UPS shipment will be billed an additional 7% of their total order **Overseas orders** $3 50 per item ordered for books sent surface mail Airmail service is available at a rate of $7 00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse Credit Card orders are accepted for all customers

Qty.	Catalog Code	Price	Amount
______	JULIA/JR	______	______

Name ______________________

Address ______________________

City ______________ State _____ Zip ______

Phone ______________________

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling ______

TOTAL ______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

CRYPTOLOGY

Albrecht Beutelspacher

This fascinating little book is eminently readable, and it is a great deal of fun to peruse... the book is a real treat. We need more books like this, crafted by expert hands yet crafted so that the general reader can enjoy them.

—Bulletin of The Institute of Combinatorics and Its Applications

This excellent and entertaining book is suitable for a first course in cryptology for mathematical enthusiasts. An abundance of exercises and an excellent list of related references are included.

—The Mathematics Teacher

In spite of the light-hearted style in which the book is written throughout, it is a serious—and successful—attempt to explain the basis of coding and decoding messages...I can strongly recommend this book to anyone who wants a brief but comprehensive, eminently readable, and up-to-date introduction to this increasingly popular topic.

— The Mathematical Gazette

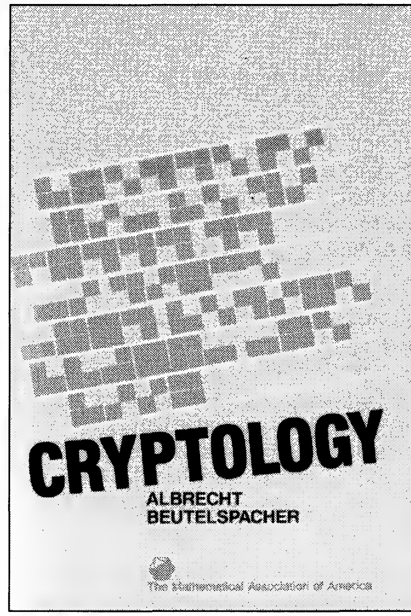
All of cryptology is covered in this work...Occupying a niche in the halls of the ivory tower of pure mathematics for nearly two millennia, number theory now forms a pillar of modern society. This book is the best explanation available today of how that pillar was constructed.

— Charles Aschbacher

A model to follow in order to make mathematics better known and understood. Accessible to a broad audience. Have fun reading this book, while you are getting a better understanding of cryptology.

— Bulletin of the Belgian Mathematics Society

How can messages be transmitted secretly? How can one guarantee that the message arrives safely in the right hands exactly as it was transmitted? Cryptology—the art and science of "secret writing"—provides ideal methods to solve these problems of data security.



The book is fun to read, and the author presents the material clearly and simply. Many exercises and references accompany each chapter.

176 pp., Paperbound, 1994
ISBN 0-88385-504-6
List: $34.00
MAA Member: $26.95
Catalog Code: CRYPT/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (202) 265-2384

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________________

Address ______________________________

City ______________________________

State ______________ Zip ______________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	CRYPT/JR	______	______
______	______________	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________________ Expires ___/___

Signature ______________________________

M THE AMERICAN MATHEMATICAL ONTHLY

Volume 104, Number 3 March 1997

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and notes should be sent to the Editor at the MONTHLY's Utah office:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send your email address and 3 copies of the complete manuscript (including all figures with captions and lettering), typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper.

Letters to the Editor on any topic are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

See the MONTHLY section of MAA Online for current information such as contents of issues, descriptive summaries of forthcoming articles, tips for authors, and preparation of manuscripts in TEX:

http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all material, typewritten on only one side of the paper.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Mr. JOE WATSON, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Some Probabilistic Aspects of Set Partitions

Jim Pitman

1. INTRODUCTION. A *partition* of the set $\mathbb{N}_n := \{1, 2, \dots, n\}$ is an unordered collection of non-empty subsets of $\mathbb{N}_n$. Let $\mathbb{P}_n$ denote the set of all such partitions, and let $B_n = \#(\mathbb{P}_n)$, the number of partitions of $\mathbb{N}_n$. The numbers B_n are known as *Bell numbers* after E.T. Bell [**3**, **4**, **45**]. See Rota [**50**] and Gardner [**24**, Chapter 2] for surveys of their properties and applications. The remarkable *Dobiński formula* [**18**]

$$B_n = e^{-1} \sum_{m=1}^{\infty} \frac{m^n}{m!} \qquad (n = 1, 2, \dots) \tag{1}$$

leads [**36**, 1.9] to the asymptotic evaluation

$$B_n \sim \frac{1}{\sqrt{n}} \lambda(n)^{n+1/2} e^{\lambda(n)-n-1} \qquad \text{as } n \to \infty, \tag{2}$$

where $\lambda(n) \log(\lambda(n)) = n$. As noted by Comtet [**11**], for each n the infinite sum in (1) can be evaluated as the least integer greater than the sum of the first $2n$ terms.

From a probabilistic perspective, the series on the right side of Dobiński's formula represents the nth moment of the Poisson distribution with mean 1. So the initially surprising fact that this series yields an integer for all n amounts to the fact that all positive integer moments of the Poisson(1) distribution are integers. As explained in Section 2, Dobiński's formula reduces to the fact that the factorial moments of the Poisson(1) distribution are identically equal to 1, and this identity can be understood probabilistically with essentially no calculation.

While such probabilistic interpretations of identities related to set partitions are the main theme of this paper, Section 1.2 recalls an elementary combinatorial proof of Dobiński's formula.

1.1 Notation. Following the notation of [**27**], let $\left\{ {n \atop k} \right\}$ denote the number of partitions of $\mathbb{N}_n$ into exactly k distinct non-empty subsets, so that

$$B_n = \sum_{k=1}^{n} \left\{ {n \atop k} \right\}. \tag{3}$$

The $\left\{ {n \atop k} \right\}$ are known as the *Stirling numbers of the second kind*. Let $m^{\underline{k}}$ denote the falling factorial with k factors

$$m^{\underline{k}} = m(m-1) \cdots (m-k+1), \tag{4}$$

which, for positive integers m and k, is the number of permutations of length k of

m distinct symbols. The formula

$$m^n = \sum_{k=1}^{n} \left\{ {n \atop k} \right\} m^{\underline{k}} \tag{5}$$

decomposes the number m^n of sequences of m distinct symbols of length n as the sum over k of the number of such sequences that contain exactly k distinct symbols [**54**, p. 35]. As an identity of polynomials in m of degree n, this identity provides an alternative definition of the coefficients $\left\{ {n \atop k} \right\}$ for $1 \le k \le n$. See [**11**, **47**, **48**, **54**] for background and a wealth of further information about Stirling numbers.

1.2 A quick proof of Dobiński's formula. This argument is attributed to Schützenberger by Foata [**22**, p. 73]. Divide (5) by $m!$ to obtain for positive integers m and n

$$\frac{m^n}{m!} = \sum_{k=1}^{m} \left\{ {n \atop k} \right\} \frac{1}{(m-k)!}. \tag{6}$$

This is the identity of coefficients of λ^m in the power series identity

$$\sum_{m=1}^{\infty} \frac{m^n}{m!} \lambda^m = \left(\sum_{k=1}^{n} \left\{ {n \atop k} \right\} \lambda^k \right) \left(\sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \right), \tag{7}$$

which, upon rearrangement, gives the following *horizontal generating function* for the Stirling numbers of the second kind:

$$\sum_{k=1}^{n} \left\{ {n \atop k} \right\} \lambda^k = e^{-\lambda} \sum_{m=1}^{\infty} \frac{m^n}{m!} \lambda^m. \tag{8}$$

Now take $\lambda = 1$ and use (3) to obtain Dobiński's formula (1). The polynomial appearing in (8) is known as an *exponential polynomial.* Many other proofs of the generalization (8) of Dobiński's formula are known. See for instance Roman [**49**, p. 66] and Wilf [**58**, p. 106]. Closely related arguments appear in Rota [**50**], Berge [**5**, p. 44], Comtet [**11**, p. 211], Lupas [**37**], and Chen-Yeh [**10**].

2. MOMENTS. For a non-negative integer-valued random variable X with

$$P(X = m) = p_m \qquad (m = 0, 1, \dots) \tag{9}$$

and a non-negative function f, let

$$E[f(X)] := \sum_m p_m f(m), \tag{10}$$

which is the *expected value of* $f(X)$ for X with distribution (9). See [**20**, **43**] for background. From (5) and linearity of the expectation operator E, we obtain the following well-known formula for $E[X^n]$, the nth *moment of* X, in terms of $E[X^{\underline{k}}]$, the kth *factorial moment* of X for $1 \le k \le n$ [**47**, **14**]:

$$E[X^n] = \sum_{k=1}^{n} \left\{ {n \atop k} \right\} E[X^{\underline{k}}]. \tag{11}$$

For $\lambda > 0$, let X_λ denote a random variable with the Poisson distribution

$$P(X_\lambda = m) = e^{-\lambda} \frac{\lambda^m}{m!} \qquad (m = 0, 1, \dots) \tag{12}$$

so that

$$E[f(X_\lambda)] = e^{-\lambda} \sum_{m=0}^{\infty} f(m) \frac{\lambda^m}{m!}. \tag{13}$$

Take $f(m) = m^n$ to see that the right side of (8) equals $E[X_\lambda^n]$, so the identity (8) amounts to the formula

$$E[X_\lambda^n] = \sum_{k=1}^{n} \left\{ {n \atop k} \right\} \lambda^k \qquad (n = 1, 2, \dots) \tag{14}$$

for the moments of the Poisson(λ) distribution **[46, 42]**. This formula is the particular case of (11) for X with Poisson(λ) distribution, for it is known **[46, 14]** that

$$E\left[X_\lambda^{\underline{k}}\right] = \lambda^k \qquad (k = 1, 2, \dots). \tag{15}$$

Formula (15) follows easily from (13) with $f(m) = m^{\underline{k}}$ by change of summation variable from m to $j = m - k$. In particular, for $\lambda = 1$ the factorial moments of the Poisson(1) distribution are identically equal to 1. So Dobiński's formula (1) can be read from (14) for $\lambda = 1$, which follows as indicated above from (11) and (15). In essence, this is Rota's **[50]** proof of Dobiński's formula cast in probabilistic notation. This argument differs from the proof in Section 1.2 in that it involves checking (15) for $\lambda = 1$.

Formula (15) has the following interpretation in terms of a *Poisson process* **[33, 43]**. Let

$$0 < U_{(1)} < \cdots < U_{(X_\lambda)} < 1 \tag{16}$$

denote the random locations in $(0, 1)$ of the points of a homogeneous Poisson process on $(0, 1)$ with mean intensity measure $\lambda\, du$ for $0 < u < 1$. For each $k = 1, 2, \dots$ define an associated *k-tuple point process*, with points in $(0, 1)^k$, to have a point at each of the locations $(U_{(\sigma_1)}, \dots, U_{(\sigma_k)})$ as σ ranges over the $X_\lambda^{\underline{k}}$ different permutations of $\{1, \dots, X_\lambda\}$ of length k. For distinct $u_i \in (0, 1)$, independence properties of the basic Poisson process on $(0, 1)$ imply that the mean intensity of the k-tuple point process at $(u_1, \dots, u_k) \in (0, 1)^k$ is

$$\frac{P\big(\text{some } U_{(\sigma_i)} \in du_i \text{ for each } 1 \le i \le k\big)}{du_1 \cdots du_k} = \frac{(\lambda\, du_1) \cdots (\lambda\, du_k)}{du_1 \cdots du_k} = \lambda^k, \tag{17}$$

so the expected number of points in the k-tuple point process is

$$E\left[X_\lambda^{\underline{k}}\right] = \lambda^k \int_0^1 du_1 \cdots \int_0^1 du_k = \lambda^k. \tag{18}$$

Constantine and Savits **[12]** derive a generalization of Dobiński's formula by consideration of compound nonhomogeneous Poisson processes. See also Stam **[52]** and Di Bucchianico **[8]** for related results. For various applications of Stirling numbers and their generalizations to the computation of moments of probability distributions, see **[47, 9]**. Moments of the normal distribution also have interesting combinatorial interpretations **[19, 25]**. More generally, the idea of representing combinatorially defined numbers by an infinite sum or an integral, typically with a probabilistic interpretation, has proved to be a very fruitful one. Other examples are the representation of $n!$ as a gamma integral, which leads to Stirling's formula **[7, 16, 38]**, and Laplace's representation of kth differences of powers **[35, 14, 30]**, which yields an asymptotic formula for the Stirling numbers of the second kind. See **[41]** for a recent survey of asymptotic enumeration methods.

3. VARIATIONS OF DOBIŃSKI'S FORMULA. The derivation of Dobiński's formula given in the previous section yields the following proposition:

Proposition 1. *Let X be a random variable with values in $\{0, 1, 2, \dots\}$ and let n be a positive integer. The following two conditions are equivalent:*

(i) *the first n factorial moments of X are identically equal to* 1;
(ii) *the kth moment of X equals B_k for every $1 \leq k \leq n$.*

It is well known that for each $\lambda > 0$ the Poisson(λ) distribution is uniquely determined by its moments; see for instance [**6**, Section 30]. The Poisson(1) distribution is therefore the unique probability distribution whose nth moment equals B_n for every n. But for each fixed n there are many probability distributions on $\{0, 1, 2, \dots\}$ that have the same first n moments as Poisson(1). It is obvious that there can be at most one such distribution of X with $P(X \leq n) = 1$, because the moment conditions amount to a system of n linearly independent equations in n unknowns $p_1, \dots, p_n$. Less obvious is the fact that the unique solution of these equations is such that $p_i \geq 0$ for $1 \leq i \leq n$ and $\sum_{i=1}^n p_i \leq 1$, so that $(p_1, \dots, p_n)$ is the restriction to $\{1, \dots, n\}$ of a unique probability distribution on $\{0, 1, \dots, n\}$. But this probability distribution on $\{0, 1, 2, \dots, n\}$, whose first n factorial moments are identically equal to one, is known to arise in the setting of the classical matching problem [**31, 14, 20, 56**]. If M_n is the number of fixed points of a uniformly distributed random permutation of $\mathbb{N}_n$, then it is easy to show by the method of indicators that the first n factorial moments of M_n are identically equal to 1; see [**14**]. The distribution of a random variable X with range $\{0, 1, \dots, n\}$ is recovered from its factorial moments by the classical sieve formula [**14**]

$$P(X = m) = \frac{1}{m!} \sum_{k=m}^{n} \frac{(-1)^{m-k} E[X^{\underline{k}}]}{(m-k)!} \quad (m = 0, 1, \dots, n). \tag{19}$$

For $X = M_n$ with $E[M_n^{\underline{k}}] \equiv 1$ for $0 \leq k \leq n$, this simplifies to

$$P(M_n = m) = \frac{1}{m!} \sum_{s=0}^{n-m} \frac{(-1)^s}{s!} \quad (m = 0, 1, \dots, n). \tag{20}$$

See [**20**, Section IV.4] for further discussion. According to Proposition 1, the kth moment of M_n equals B_k for every $1 \leq k \leq n$. That is to say,

$$B_k = \sum_{m=1}^{n} \frac{m^k}{m!} \sum_{s=0}^{n-m} \frac{(-1)^s}{s!} \quad (1 \leq k \leq n). \tag{21}$$

This variation of Dobiński's formula is derived in quite a different way by Wilf [**58**, p. 22] by substituting the classical formula

$$\left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{j=1}^{k} (-1)^{k-j} \binom{k}{j} j^n \tag{22}$$

into (3). As observed by Wilf, Dobiński's formula (1) follows easily from (21) by letting $n \to \infty$. See also Lovász [**36**, 1.9] for a similar argument, James and Kerber [**32**, pp. 227–237] for connections with the representation theory of the symmetric group, and Diaconis and Shashahani [**17**] for various generalizations. In Dale and Skau [**13**] the Bell numbers appear as the factorial moments of a probability distribution on the non-negative integers.

4. THE MOMENT GENERATING FUNCTION. Consider now the *moment generating function* (*m.g.f.*) of the Poisson(λ) distribution:

$$E[\exp(\theta X_\lambda)] = E\left[\sum_{n=0}^{\infty} \frac{\theta^n X_\lambda^n}{n!}\right] = \sum_{n=0}^{\infty} E[X_\lambda^n]\frac{\theta^n}{n!}, \tag{23}$$

where the series converge for all real θ and the interchange of E and Σ is easily justified. See **[6]** for a modern treatment of m.g.f's. From (13) with $f(m) = e^{\theta m}$ there is the standard formula

$$E[\exp(\theta X_\lambda)] = e^{-\lambda} \sum_{m=0}^{\infty} \frac{(\lambda e^\theta)^m}{m!} = \exp(\lambda(e^\theta - 1)). \tag{24}$$

This combines with (8) to yield the following *double generating function* of the Stirling numbers of the second kind. This classical formula **[11**, p. 206] is an identity between two different expressions for the m.g.f. in θ of the Poisson(λ) distribution:

$$1 + \sum_{n=1}^{\infty} \sum_{k=1}^{n} \left\{ {n \atop k} \right\} \frac{\lambda^k \theta^n}{n!} = \exp(\lambda(e^\theta - 1)). \tag{25}$$

In particular, for $\lambda = 1$ this reduces by (3) to Bell's **[3, 4]** formula

$$1 + \sum_{n=1}^{\infty} B_n \frac{\theta^n}{n!} = \exp(e^\theta - 1), \tag{26}$$

which gives two expressions for the m.g.f. in θ of the Poisson(1) distribution. Equating coefficients of λ^k in (25) yields the *vertical generating function* of the Stirling numbers of the second kind:

$$\sum_{n \geq k} \left\{ {n \atop k} \right\} \frac{\theta^n}{n!} = \frac{1}{k!}(e^\theta - 1)^k \qquad (k = 1, 2, \dots). \tag{27}$$

See **[11, 47, 54]** for alternative derivations of these identities. There are similar identities for many other arrays of combinatorial numbers, such as the binomial coefficients and Stirling numbers of the first kind **[11, 58]**, **[27**, p. 351], most of which admit probabilistic interpretations. Formulae with binomial coefficients typically involve independent trials, while those with Stirling numbers of the first kind typically involve the cycle structure of random permutations **[1]**. See also **[2]** for probabilistic analysis of more general combinatorial structures and further references.

5. RANDOM PARTITIONS. A *random partition of* $\mathbb{N}_n$ is a random variable Π with values in the set $\mathbb{P}_n$ of partitions of $\mathbb{N}_n$. The *distribution of* Π then refers to the collection of probabilities $P(\Pi = \pi)$ as π ranges over $\mathbb{P}_n$. Questions about enumeration of partitions of $\mathbb{N}_n$ of various kinds can be phrased probabilistically in terms of a *uniform random partition*, that is, a random partition Π with the uniform distribution $P(\Pi = \pi) = 1/B_n$ for each partition $\pi \in \mathbb{P}_n$. For developments of this idea see **[29, 28, 51, 23]**. Random partitions with non-uniform distribution also arise naturally in various contexts, so it is useful to have models for random partitions, both uniform and non-uniform.

The following *random allocation scheme* provides a basic method of generating a random partition of $\mathbb{N}_n$. See **[14, 34, 57]** for extensive study of this and related schemes, and further references. Throw n balls labelled by $\mathbb{N}_n$ into m boxes

labelled by $\mathbb{N}_m$, and assume that all m^n possible allocations of balls into boxes are equally likely. Let Π_{nm} be the partition of balls by boxes. More formally, let X_i be the number of the box containing the ith ball for $1 \le i \le n$. Then the X_i are independent and uniformly distributed on $\mathbb{N}_m$, and Π_{nm} is the partition of $\mathbb{N}_n$ induced by the random equivalence relation $i \sim j$ if and only if $X_i = X_j$. Formally, the X_i can be regarded as coordinate maps defined on $(\mathbb{N}_n)^m$, and Π_{nm} is then defined as a map from $(\mathbb{N}_n)^m$ to $\mathbb{P}_n$, the set of partitions of $\mathbb{N}_n$. Let $\#(\pi)$ denote the number of subsets in a partition $\pi \in \mathbb{P}_n$. The distribution of Π_{nm} induced by the uniform distribution P on $\mathbb{N}_m$ can be read from formula (5):

$$P(\Pi_{nm} = \pi) = \frac{m^{\underline{k}}}{m^n} \qquad \text{if } \#(\pi) = k. \tag{28}$$

The distribution of $\#(\Pi_{nm})$, the number of occupied boxes when n balls are thrown into m boxes, is given by the following probabilistic equivalent of (5):

$$P[\#(\Pi_{nm}) = k] = \left\{ {n \atop k} \right\} \frac{m^{\underline{k}}}{m^n} \qquad (1 \le k \le n). \tag{29}$$

Because the probability displayed in (28) depends on the number of occupied boxes k, for $n \ge 3$ this random partition Π of $\mathbb{N}_n$ does not have uniform distribution on $\mathbb{P}_n$ for any m. However, as observed by Stam **[53]**, for each fixed n it is possible to generate a uniformly distributed random element of $\mathbb{P}_n$ by a suitable randomization of m. The following proposition was suggested by Stam's construction, which is described in Corollary 3.

Proposition 2. *Let M be a random variable with values in* $\{1, 2, \dots\}$, *and suppose given $M = m$ that n balls labelled by $\mathbb{N}_n$ are thrown independently and uniformly at random into m boxes. Let Π_{nM} denote the random partition of $\mathbb{N}_n$ so generated. The following two conditions are equivalent*:

(i) *Π_{nM} has uniform distribution over the set $\mathbb{P}_n$ of all partitions of $\mathbb{N}_n$*;
(ii) *the distribution of M is of the form*

$$P(M = m) = \frac{m^n p_m}{B_n} \qquad (m = 1, 2, \dots) \tag{30}$$

for some probability distribution $(\mathrm{p_m})$ *on* $\{0, 1, 2, \dots\}$ *whose first n factorial moments are identically equal to* 1.

Before the proof, here are two corollaries which follow immediately from the Proposition and the discussion in Sections 2 and 3:

Corollary 3. **[53]** *If M has the distribution* (30) *for $p_m = e^{-1}/m!$, then Π_{nM} has uniform distribution on $\mathbb{P}_n$.*

Corollary 4. *For each n there is a unique distribution of M such that*

$$P(M \le n) = 1 \textit{ and } \Pi_{nM} \textit{ has uniform distribution on } \mathbb{P}_n.$$

This distribution of M is defined by (30) *for $p_m = P(M_n = m)$ as in* (20) *with M_n the number of fixed points of a uniform random permutation of $\mathbb{N}_n$.*

Proof of Proposition 2. By conditioning on M and using (28),

$$P(\Pi_{nM} = \pi) = \sum_{m=1}^{\infty} \frac{m^{\underline{k}}}{m^n} P(M = m) \qquad \text{if } \#(\pi) = k, \tag{31}$$

so the distribution of Π_{nM} is uniform on $\mathbb{P}_n$ if and only if

$$\sum_{m=1}^{\infty} \frac{m^{\underline{k}}}{m^n} P(M = m) = \frac{1}{B_n} \qquad (1 \le k \le n). \tag{32}$$

Define

$$p_m = B_n m^{-n} P(M = m) \qquad (m = 1, 2, \dots), \tag{33}$$

so that (32) becomes

$$\sum_{m=1}^{\infty} m^{\underline{k}} p_m = 1 \qquad (1 \le k \le n), \tag{34}$$

which for $k = 1$ implies that $\sum_{m=1}^{\infty} p_m \le \sum_{m=1}^{\infty} m p_m = 1$. It follows that Π_{nM} is uniform if and only if (p_m) derived from the distribution of M via (33) is the restriction to $\{1, 2, \dots\}$ of a probability distribution on $\{0, 1, 2, \dots\}$ whose first n factorial moments are equal to 1. This is condition (ii). ∎

As shown by Stam, Corollary 3 allows numerous results regarding the asymptotic distribution for large n of a uniform random partition of $\mathbb{N}_n$ to be deduced from corresponding results for the classical occupancy problem defined by random allocations of balls in boxes, for which see **[34, 57]**. See also **[2, 15, 23, 26, 28, 29, 51]** for a more detailed account of the asymptotics of uniform random partitions of $\mathbb{N}_n$.

As a variation, the following corollary is easily obtained by a similar argument:

Corollary 5. *Suppose that M has the distribution*

$$P(M = m) = \frac{m^n P(X_\lambda = m)}{\mu_n(\lambda)} \qquad (m = 1, 2, \dots), \tag{35}$$

where X_λ has the Poisson(λ) distribution (12), *and $\mu_n(\lambda) = E(X_\lambda^n)$. Then the distribution of Π_{nM} is given by*

$$P(\Pi_{nM} = \pi) = \frac{\lambda^k}{\mu_n(\lambda)} \qquad \text{if } \#(\pi) = k. \tag{36}$$

As a check, (36) implies

$$P[\#(\Pi_{nM}) = k] = \left\{ {n \atop k} \right\} \frac{\lambda^k}{\mu_n(\lambda)} \qquad (1 \le k \le n). \tag{37}$$

The fact that these probabilities sum to 1 amounts to formula (14) for $\mu_n(\lambda)$. The distribution of Π_{nM} defined by formula (36) defines a *Gibbs distribution* on partitions of $\mathbb{N}_n$. See **[55, 44]** for further discussion of such Gibbs distributions on sets of combinatorial objects. See Nijenhuis and Wilf **[39]** for a recursive algorithm to construct a uniform random partition of $\mathbb{N}_n$ based on the recurrence

$$B_n = 1 + \sum_{k=1}^{n-1} \binom{n-1}{k} B_k, \tag{38}$$

where the right side counts the number of partitions π of $\mathbb{N}_n$ according to the size k of the subset in π that contains n **[36**, Problem 1.10**]**. See **[40]** for related

combinatorial algorithms, and **[21]** for a recent systematic approach to the random generation of labelled combinatorial structures and further references on this topic.

ACKNOWLEDGMENTS. Thanks to Persi Diaconis, Hsien-Kuei Hwang, and the referee for providing some references.

REFERENCES

1. R. Arratia and S. Tavaré. The cycle structure of random permutations, *Ann. Prob.*, 20:1567–1591, 1992.
2. R. Arratia and S. Tavaré. Independent process approximations for random combinatorial structures, *Adv. Math.*, 104:90–154, 1994.
3. E. T. Bell. Exponential numbers, *Amer. Math. Monthly*, 41:411–419, 1934.
4. E. T. Bell. Exponential polynomials, *Annals of Mathematics*, 35:258–277, 1934.
5. C. Berge. *Principles of Combinatorics*. Academic Press, New York and London, 1971.
6. P. Billingsley. *Probability and Measure*, 2*nd ed.* Wiley, New York, 1986.
7. C. R. Blyth and P. K. Pathak. A note on easy proofs of Stirling's theorem, *Amer. Math. Monthly*, 93:376–379, 1986.
8. A. Di Bucchianico. Representations of Sheffer polynomials, *Studies in Applied Math.*, 93:1–14, 1994.
9. Ch. A. Charalambides and J. Singh. A review of the Stirling numbers, their generalizations and statistical applications, *Commun. Statist.-Theory Meth.*, 17:2533–2595, 1988.
10. B. Chen and Y.-N. Yeh. Some explanations of Dobinski's formula, *Studies in Applied Math.*, 92:191–199, 1994.
11. L. Comtet. *Advanced Combinatorics*. D. Reidel Pub. Co., Boston, 1974. (translated from French).
12. G. M. Constantine and T. H. Savits. A stochastic representation of partition identities, *SIAM J. Discrete Math.*, 7:194–202, 1994.
13. K. Dale and I. Skau. The (generalized) secretary's packet problem and the Bell numbers, *Discrete Math.*, 137:357–360, 1995.
14. F. N. David and D. E. Barton. *Combinatorial Chance*. Griffins, London, 1962.
15. J. M. DeLaurentis and B. G. Pittel. Counting subsets of the random partition and the 'Brownian bridge' process, *Stochastic Processes and their Applications*, 15:155–167, 1983.
16. P. Diaconis and D. Freedman. An elementary proof of Stirling's formula, *Amer. Math. Monthly*, 93:123–125, 1986.
17. P. Diaconis and M. Shahshahani. On the eigenvalues of random matrices, *J. Applied Probability*, Special vol. 31A:49–62, 1994.
18. G. Dobiński. Summirung der Reihe $\Sigma n^m/n!$ für $m = 1, 2, 3, 4, 5, \dots$, *Grunert Archiv* (*Arch. für Mat. und Physik*), 61:333–336, 1877.
19. E. B. Dynkin. Gaussian and non-Gaussian random fields associated with Markov processes, *J. Funct. Anal.*, 55:344–376, 1984.
20. W. Feller. *An Introduction to Probability Theory and its Applications*, Vol. 1, 3rd ed. Wiley, New York, 1968.
21. Ph. Flajolet, P. Zimmerman, and B. V. Cutsem. A calculus for the random generation of labelled combinatorial structures, *Theoretical Computer Science*, 132:1–35, 1994.
22. D. Foata. *La série génératrice exponentielle dans les problèmes d'énumération*, volume 54 of *Séminaire de Mathématiques supérieures*. Presses de l'Université de Montréal, 1974.
23. B. Fristedt. The structure of partitions of large sets, Technical Report 86-154, University of Minnesota Mathematics Dept., Minneapolis, Minnesota, 1987.
24. M. Gardner. *Fractal music, hypercards and more— : mathematical recreations from Scientific American magazine*. W.H. Freeman, New York, 1992.
25. C. D. Godsil. *Algebraic Combinatorics*. Chapman & Hall, New York, 1993.
26. W. M. Y. Goh and E. Schmutz. Random set partitions, *SIAM J. Discrete Math.*, 7:419–436, 1994.
27. R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*: *a foundation for computer science*, 2*nd ed.* Addison-Wesley, Reading, Mass., 1989.
28. J. Haigh. Random equivalence relations, *Journal of Combinatorial Theory*, 13:287–295, 1972.
29. L. H. Harper. Stirling behavior is asymptotically normal, *Ann. Math. Stat.*, 38:410–414, 1966.
30. L. Holst. On numbers related to partitions of unlike objects and occupancy problems, *Europ. J. Combinatorics*, 2:231–237, 1981.

31. J. O. Irwin. A unified derivation of some well-known frequency distribution of interest in biometry and statistics, *J.R. Statist. Soc.*, 118:389–404, 1955.
32. G. D. James and A. Kerber. *The representation theory of the symmetric group*, volume 16 of *Encyclopedia of mathematics and its applications*. Addison-Wesley, Reading, Mass., 1981.
33. J. F. C. Kingman. *Poisson Processes*. Clarendon Press, Oxford, 1993.
34. V. F. Kolchin, B. A. Sevastyanov, and V. P. Christyakov. *Random Allocations*. V. H. Winston & Sons, Washington, D. C., 1978.
35. P. S. Laplace. *Théorie Analytique des Probabilités, 2nd ed.*. Paris, 1814.
36. L. Lovász. *Combinatorial Problems and Exercises, 2nd ed.* North-Holland, Amsterdam, 1993.
37. A. Lupas. Dobiński-type formula for binomial polynomials, *Stud. Univ. Babes-Bolyai Math.*, 33(2):40–44, 1988.
38. V. Namias. A simple derivation of Stirling's asymptotic series, *Amer. Math. Monthly*, 93:25–29, 1986.
39. A. Nijenhuis and H. S. Wilf. A method and two algorithms on the theory of partitions, *J. Combin. Theory A*, 18:219–222, 1975.
40. A. Nijenhuis and H. S. Wilf. *Combinatorial algorithms for computers and calculators, 2nd ed.* Academic Press, New York, 1978.
41. A. M. Odlyzko. Asymptotic enumeration methods, in R. L. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of Combinatorics Vol. II*, pp. 1063–1229. Elsevier, New York, 1995.
42. C. Philipson. A note on the moments of a Poisson probability distribution, *Skand. Aktuarietidskrift*, 46:243–4, 1963.
43. J. Pitman. *Probability*. Springer-Verlag, New York, 1993.
44. J. Pitman. Probabilistic bounds on the coefficients of polynomials with only real zeros, Technical Report 453, Dept. Statistics, U.C. Berkeley, 1996. To appear in *J. Comb. Theory A*.
45. C. Reid. *The search for E. T. Bell: also known as John Taine*. MAA Spectrum. Mathematical Association of America, Washington, DC, 1993.
46. J. Riordan. Moment recurrence relations for the binomial Poisson and hypergeometric probability distributions, *Ann. Math. Stat.*, 8:103–111, 1937.
47. J. Riordan. *An Introduction to Combinatorial Analysis*. Wiley, New York, 1958.
48. J. Riordan. *Combinatorial Identities*. Wiley, New York, 1968.
49. S. Roman. *The Umbral Calculus*. Academic Press, New York, 1984.
50. G.-C. Rota. The number of partitions of a set, *Amer. Math. Monthly*, 71:498–504, 1964.
51. V. N. Sachkov. Random partitions of sets, *Theory Probab. Appl.*, 19:184–190, 1973.
52. A. J. Stam. Polynomials of binomial type and compound Poisson processes, *Journal of Mathematical Analysis and Applications*, 130:151, 1988.
53. A. J. Stam. Generation of a random partition of a set by an urn model, *J. Combin. Theory A*, 35:231–240, 1983.
54. R. P. Stanley. *Enumerative Combinatorics*, Vol I. Wadsworth & Brooks/Cole, Monterey, California, 1986.
55. J. M. Steele. Gibbs' measures on combinatorial objects and the central limit theorem for an exponential family of random trees, *Probability in the Engineering and Informational Sciences*, 1:47–59, 1987.
56. L. Takács. The problem of coincidences, *Arch. Hist. Exact. Sci.*, 21:229–244, 1980.
57. V. A. Vatutin and V. G. Mikhailov. Limit theorems for the number of empty cells in an equiprobable scheme for group allocation of particles, *Theory of Probability and its Applications* (*Transl. of Teorija Verojatnostei i ee Primenenija*), 27:734–743, 1982.
58. H. Wilf. *Generatingfunctionology*. Academic Press, San Diego, 1990.

Department of Statistics
University of California
367 Evans Hall # 3860
Berkeley, CA 94720-3860
pitman@stat.Berkeley.EDU

Building an International Reputation: The Case of J. J. Sylvester (1814–1897)

Karen Hunger Parshall and Eugene Seneta

James Joseph Sylvester—prolific, gifted, flamboyant, egocentric, cantankerous. At the time of his death in London on 15 March, 1897, Sylvester's reputation internationally as one of the nineteenth century's principal mathematical figures had long been secure. He had worked hard to assure this. Obviously, he had done much seminal work in building the theory of invariants, and this had contributed to his renown. Yet, Sylvester had felt compelled to establish ties directly with mathematicians at home—but more importantly abroad—in order to make his name known. Was this just a matter of egocentrism, or did other factors contribute to his international focus? What did it take to become an internationally recognized British mathematician in the latter half of the nineteenth century, when first France and then Germany very much set the mathematical standard? Why was this even important? As the centenary of Sylvester's death brings historians and mathematicians together in England to celebrate his life and research, we examine some of the reasons why Sylvester valued the international mathematical arena so highly and how he used it to his advantage during the course of his career.

It is well-known that Sylvester, as a Jew, was, like all non-Anglicans, debarred by the Test Acts from taking the Cambridge degree he had earned as Second Wrangler in 1837 and from holding a Cambridge fellowship or professorship. His first position, the professorship of natural philosophy at nonsectarian University College London, was too far from his real interests and expertise to satisfy him, so he gave it up in 1841 after only three years for the uncertain fortunes of a professorship of mathematics in exile far from home (and, he quickly came to think, far from civilization!) at the University of Virginia. He lasted there for four-and-a-half months before resigning over a matter of principle and fleeing northward to New York City and his brother's home. From there, he tried in vain for some eighteen months to secure a new position in the United States—at Columbia College, at Harvard with Benjamin Peirce, in the Washington, D.C. area, and even at the University of South Carolina—before returning to England to resume what he termed the "fruitless and hopeless struggle with an adverce [sic] tide of affairs" **[39]**. By the close of 1844, though, he proudly reported having "recovered [his] footing in the world's slippery path" **[39]** thanks to his assumption of the post of actuary and secretary at the Equity and Law Life Assurance Company in London. During the decade from 1845 to 1855, he prepared for and passed the Bar; he met his mathematical alter ego, Arthur Cayley; and he produced his ground-breaking series of papers in what would come to be known as invariant theory. The next fifteen years found him in his first sustained academic post, the professorship of mathematics at the Royal Military Academy in Woolwich, where he taught drudgerous mathematics to mostly uncaring students and fought with the military authorities over teaching loads destined, he was convinced, to bring the "extinction of my scientific existence" **[40]**. Sylvester's career trajectory

James Joseph Sylvester, aged 26, by George Patton. In the private collection of Mr. and Mrs. Alain Enthoven.

clearly diverged from those guided within the ivy-covered walls of the colleges of Cambridge and Oxford.

As an establishment-outsider, who, unlike Augustus DeMorgan, did not even hold a position at the leading anti-establishment institution, how was Sylvester to secure the reputation that his healthy-sized ego demanded and that his manifest mathematical talents warranted? Surely, the British Association for the Advancement of Science or the Royal Society, to which he had been elected in 1839, represented avenues toward the establishment of national recognition, and Sylvester presented his work—like accounts of his research on Sturm's theorem for locating the roots of an algebraic equation ([**1**, 1:59–60] and [**1**, 1:429–586])—be-

fore both of these organizations throughout his career. In Sylvester's Platonistic view, however, mathematics was a universal endeavor that transcended national boundaries (see **[24]** for an analysis). Moreover, as Sylvester well realized, the Continent—and particularly France and Germany—dominated mathematical research at mid-century, while England tended to assume a more isolationist posture. It was thus important to make one's work known abroad. This would help assure that credit was given where due, that results published in British journals were not ignored or overlooked outside England, and that British mathematicians effectively contributed to building the eternal edifice of mathematics. As Marin Mersenne had shown in the seventeenth century, establishing an international network of correspondents could be remarkably helpful in achieving these goals.

Sylvester actively sought to forge his own international mathematical connections beginning around 1850. Initially, at least, he seemed to have the greatest number of ties with France, a country which had dominated mathematics and the sciences during the first half of the nineteenth century **[14]**, a country with whose language Sylvester felt at ease, a country with an influential scientific society and mathematicians of the highest repute. In France, Sylvester enjoyed perhaps his most lasting and most intimate mathematical association with Charles Hermite.

Eight years Sylvester's junior, Hermite had come upon the mathematical scene in the 1840s with work on elliptic and hyperelliptic functions that had earned Jacobi's admiration. In 1848, the year after taking his *baccalauréat* and *licence*, he was named *répétiteur* and admissions examiner at the *École polytechnique*, where he himself had pursued his studies. Rather quickly, he established a reputation as one of the rising stars in French mathematics with his work on quadratic forms and, beginning in 1854, on the theory of invariants. His election to the Paris *Académie des Sciences* in 1856 gives a clear indication of his stature in the French scientific community. Since Hermite's research interests fundamentally overlapped those of Sylvester, it is little wonder that the Englishman sought out the kindred French mathematical spirit. In particular, Sylvester wanted to make sure that Hermite knew of and gave the proper credit to his research.

Charles Hermite, aged 25, engraved by Ch. Wittmann from *Oeuvres de Charles Hermite*, ed. Émile Picard, 3 vols. (Paris: Gauthier-Villars, 1905–1912).

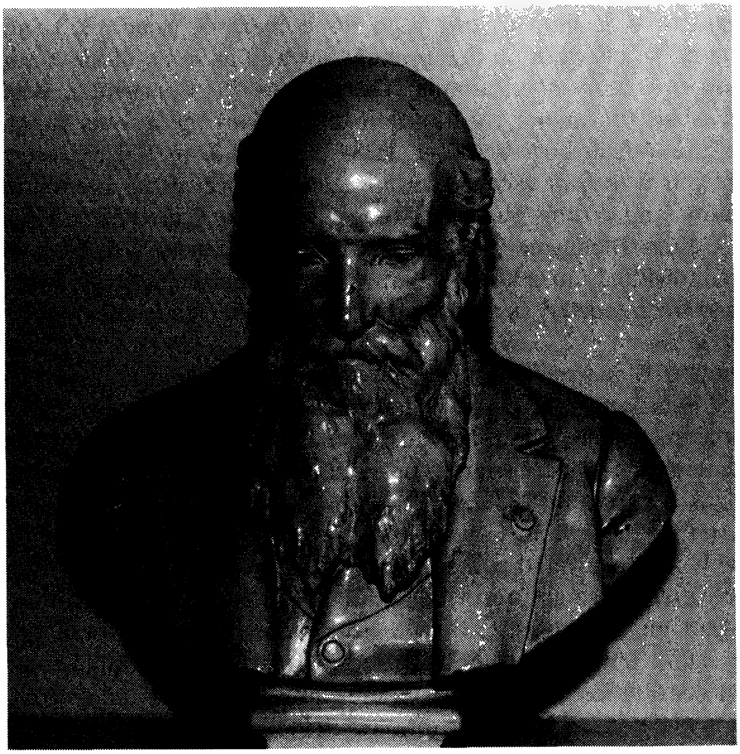
After finishing his law studies in 1850 and following the establishment of his close personal friendship and mathematical exchange with Arthur Cayley, Sylvester finally began to come into his own as a researcher. Between 1850 and 1854, he published much of his work on determinants as well as some of his seminal papers on the emergent theory of invariants. The year 1852, in particular, witnessed the publication of his first major contribution to invariant theory, his paper, "On the Principle of the Calculus of Forms" [**1**, 1:284–327 and 328–363]. Sylvester earnestly desired that this work *not* escape the immediate notice of the French mathematical community, and, to this end, he wrote several letters in 1852 to his correspondent, Irenée-Jules Bienaymé (compare [**20**]), asking him to distribute offprints to certain key individuals in addition to the *Société philomatique de Paris*, to which Sylvester had been elected as a corresponding member early that year, and to the *Institut de France*.

According to Bienaymé's handwritten tally [**2**] in response to Sylvester's first request on 7 February, 1852 [**31**], Michel Chasles, Charles Hermite, Olry Terquem, Eugène Catalan, Joseph Serret, and Joseph Bertrand were to receive copies of the first installment of Sylvester's latest pronouncement on the calculus of forms [**1**, 1:284–327] as well as an 1851 paper, "On the General Theory of Associated Forms" [**1**, 1:198–202], while Cauchy and the *Institut de France* were to get these together with a paper on the theory of determinants [**1**, 1:241–250]. Bienaymé dutifully delivered these and received additional requests from Sylvester on 4 June [**32**] and again on 27 August [**33**] to deliver more papers to these and selected others. Sylvester's letter of 4 June revealed at least one additional motivation for this general distribution of his work. There, he asked Bienaymé to pass copies of the second installment of his paper on the calculus of forms to Hermite and to the *Société philomatique* as before, but this time he also wanted a copy to go to the editor of the *Journal des mathématiques pures et appliquées*, Joseph Liouville. As he explained to Bienaymé,

> I wish M. Liouville to have a copy because I am told that M. Eisenstein of Berlin has sent to Liouville's Journal the same kind of matter as is in my Section VI [**32**].

This was the demonstration of the sufficiency of two particular differential operators that Cayley and Sylvester employed for detecting invariants [**1**, 1:351–360], and Sylvester clearly wanted Liouville to know that he and Cayley—and *not* Eisenstein—had discovered the properties of these operators first. Eisenstein did not end up publishing the contested proof [**11**].

As Sylvester's delivery list to Bienaymé also indicates, the Englishman's association with Michel Chasles dates from at least the early 1850s. Chasles, who had established his reputation as both a geometer and an historian of mathematics as early as 1837 with the publication of his *Aperçu historique sur l'origine et le développement des méthodes en géométrie*, had been named to the chair of higher geometry at the Sorbonne in 1846 and had become a full member of the Paris *Académie des Sciences* in 1851. He was thus another influential member of the French mathematical community and a worthy contact for Sylvester to establish, even though his brand of geometric research was never Sylvester's forte. In 1852, as Sylvester was busy with the early invariant-theoretic work that he was anxious for Bienaymé to deliver to Chasles, Chasles had just published a new book, his *Traité de géométrie supérieure* [**8**], and asked Sylvester to serve as the messenger for two



Irenée-Jules Bienaymé (1796–1878) by Jules Franceschi (1825–1893). Photo by, and courtesy of, Arnaud Bienaymé.

copies to British colleagues [7]. In his letter of 26 August, Chasles also provided Sylvester with an indication that he was starting to make a name for himself in France.

Chasles reported that

> I saw M. Terquem yesterday; he spoke to me with pleasure and enthusiasm of the beautiful theorem that you sent him, and knowing that I was going to write you, he asked me to give you his compliments and to tell you that he was going to publish your communication without delay in his journal [**7**, our translation].

Olry Terquem was the editor of the *Nouvelles annales de mathématiques*, and the submission in question was Sylvester's first publication in a foreign journal, his determinant-theoretic proof of the fact that if A is an $n \times n$ symmetric matrix, then the roots of the characteristic equation of A^p for p an integer are the roots of the characteristic equation of A, each raised to the pth power [**1**, 1:364–366]. Sylvester quickly followed this with two more notes to Terquem's journal ([**1**, 1:423] and [**1**, 1:424–428]). Having successfully taken the step of bringing his work directly before the Continental audience by publishing in a foreign journal, Sylvester continued throughout his career to submit his work to periodicals both at home and abroad.

Besides Hermite, Bienaymé, and Chasles, Sylvester also established ties in the 1850s in France with Jean-Victor Poncelet, Jean-Marie-Constant Duhamel, Joseph Serret, and Joseph Bertrand. Relative to Germany, however, his contacts were initially fewer and his relations never as close. One exception to this was Carl Borchardt, who had spent the year from 1846 to 1847 studying in Paris and forging his own links with Chasles, Hermite, and Liouville. Borchardt had earned his doctorate under Jacobi at the University of Königsberg in 1843 and had made a splash in 1846 with his first mathematical publication, a work right up Sylvester's mathematical alley. Borchardt had given a determinant-theoretic argument showing that the Sturm functions that arise from the equation determining the secular disturbances of the planets can be represented as the sum of squares [**4**]. In subsequent work, he considered related algebraic questions involving symmetric functions and elimination theory, that also actively engaged Sylvester. Given their common interests, it is little wonder that Sylvester and Borchardt began to correspond. It is also little wonder that, in 1852, Sylvester wanted to apprise Borchardt of his latest researches.

On 20 February, just two weeks after he had entrusted Bienaymé with the bundle of offprints for distribution in Paris, Sylvester wrote Borchardt enclosing a copy of one of the same papers, the first installment of his paper "On the Principle of the Calculus of Forms." Burdened by his ongoing preparation of Jacobi's collected works and by his teaching duties as *Privatdozent* at the University of Berlin, Borchardt only responded to Sylvester's letter on 6 April [**3**]. There, he apologized for the delay in his response and for not having had a chance to read and study Sylvester's latest achievement. More interestingly, he also returned to a topic that had apparently been under discussion in earlier letters between the two men: the integrity of Otto Hesse.

Because the early 1850s were abundantly productive years for Sylvester mathematically, it was then that he actively sought to build and solidify his increasingly deserved reputation. Always touchy about matters of priority, but perhaps most touchy during these early years, Sylvester came to feel that Hesse [18] had stolen a result he had published in the *Philosophical Magazine* in 1841 [**1**, 1:75–85]. Sylvester, furious, felt that Hesse had consciously failed to credit his work and lambasted him in print in [**1**, 1:184–197 on p. 189]. In his letter, however, Borchardt offered a very different read on the situation. In his measured and more objective view, Borchardt offered that

> [i]f Mr. Hesse had known of your memoir ..., he would not have committed plagiarism, and I know him too well to believe him capable of it. This changes nothing relative to your priority but much relative to the moral judgment on Mr. Hesse. In mathematics it often happens that, owing to an insufficient knowledge of the literature, results are published as new which have already been obtained earlier by others. Such oversights must certainly be corrected, but if in every such case of this kind one claimed plagiarism, one would not be justified [**3**, our translation].

This dispassionate assessment of the situation as well as Sylvester's heated reaction to what he deemed to be Hesse's initial slight reflect a key aspect of reputation-building, the paramount importance of priority. Without it, someone else's reputation may grow incrementally at the expense of one's own hard work and effort. In Sylvester's case, this became even more of an issue when publications serving primarily different national constituencies were involved. If Hesse read the *Philo-*

sophical Magazine, a periodical that published, by and large, mathematical work of lesser quality than its German counterparts, then he had *overlooked* Sylvester's paper. If he—and by extension his countrymen—did *not* read the *Philosophical Magazine*, then all the worse for those British mathematicians trying conscientiously to make their work known to the broader community of mathematicians through that means. At least in this case, the priority dispute was ultimately little more than a tempest in a teapot; just two years later, at a crucial juncture in Sylvester's career, Hesse served as one of the Englishman's hand-picked references.

By 1854, Sylvester had been in his job as actuary and secretary at Equity Law and Life for some nine years. He had also been doing some of his best mathematical research and, understandably, wanted a position more consonant with his interests and training. When the professorship of mathematics at the Royal Military Academy in Woolwich came open, then, he quite naturally took the opportunity to apply. When he lost out on the appointment to a mathematically inferior candidate and felt that the military authorities had misrepresented some of the facts surrounding the election, he wrote to another longtime acquaintance, the former Lord Chancellor, Lord Brougham, expressing his frustration. Not without a certain amount of pride, Sylvester let Brougham know that

> Letters were written or sent in in support of my application and couched in the strongest language in which a recommendation could be clothed from
>
> Sir William Hamilton, Dublin
> Professor Graves, D[itt]o
> Professor Kelland, Edinburgh
> Professor Challis, Cambridge
> The Bishop of Natal
> General Poncelet, Paris
> M. Chasles, Paris
> Rev[eren]d Geo Salmon, Dublin
> Duhamel, Paris
> Serret, Hermite, Bertrand } of the Examiners at the Polytechnique
>
> and many others.
>
> Letters were also written but too late to be sent in by
>
> the great Lejeune Dirichlet, Berlin
> Professors Peters & Hesse, Königsberg
> Professor Joachimsthal, Halle
> Professor Thomson, Glasgow
>
> and also from distinguished pupils testifying to my teaching powers [**34**].

As is evident from this roster, Sylvester made important use of the international network he had established, in his attempt to break into the English academic world. This first try proved unsuccessful, but, when the victorious candidate died unexpectedly after only a few months on the job, Sylvester reapplied and won the appointment (see [**28**] on the changing London mathematical scene). His distinguished list of referees may not have helped him get the post, but it apparently did not hurt. Sylvester had finally broken back into academe after a dozen years away.

The late 1850s and early 1860s found Sylvester hard at work on his ongoing researches in invariant theory and in new, but not unrelated, work in combinatorics. In a series of seven lectures delivered at King's College, London between 6 June and 11 July, 1859 (but published only in 1897), he brought together a large number of known and new combinatorial results in an early effort to systematize a *theory* of partitions [**1**, 2:119–175]. Prior to delivering the last of these lectures, Sylvester wrote to his friend, Pafnuty Chebyshev, thanking him for the elucidating remarks on some of Euler's work on partitions that enabled him to include a discussion of it in his presentation [**38**]. Sylvester and Chebyshev had known each other at least since the fall of 1852, when Chebyshev was on a foreign tour to study the scientific and technological advances in evidence in Paris, London, and Berlin. After seeing Bienaymé and others in Paris, he went on to visit Sylvester in London, and the two remained in touch sporadically for many years thereafter. In fact, like Sylvester, Chebyshev also maintained his contact with Hermite. As the Russian mathematician would later recollect,

> [w]e were once sitting in Paris, the three of us: Hermite, Sylvester, and I. Hermite–the leading mathematician of France, Sylvester—the leading mathematician of England, and I [**15**, our translation].

Sylvester's growing international network also extended to Italy by the early 1860s to include Enrico Betti, fellow determinant-theorist Francesco Brioschi, and physicist and later Minister of Education, Carlo Matteucci, among others. In the winter of 1862, in fact, Sylvester followed one his many trips to Paris with an extended scientific tour of the newly unified Italy and described at least part of his itinerary in a letter to Cayley [**36**]. Here, Sylvester entered into a different role relative to his international network; his interest—as well as that of other foreign mathematicians—in the post-unification Italian scene helped to legitimize the research efforts of the Italians in the latter half of the nineteenth century ([**5**] and [**22**]). Sylvester's presence helped the Italians establish *their* reputations; by the early 1860s, *his* reputation, at least in their eyes, was already secure. This was apparently true elsewhere as well.

Late in 1863, Sylvester received what must have been the most significant symbol of his international stature as a mathematician, his election as foreign correspondent to the Paris *Académie des Sciences*. It was his contact, Joseph Bertrand, whom Sylvester chose to communicate one of his first "official" works to the assembled *savants*, and that communication related to his (the first) proof of Newton's rule for isolating pairs of complex roots of polynomial equations [**1**, 2:361–362]. As Sylvester explained to Bertrand,

> I have proved, not without some difficulty, Newton's rule to the fifth degree inclusive. Messers De Morgan and Cayley have expressed to me their firm conviction that Newton himself never had a proof of this rule, which remains to this day the marvel and the disgrace of Algebra [**30**, our translation].

Interestingly, Sylvester announced the result in print first in France, but only after having read the paper before the Royal Society; he later published the full exposition of it in the *Philosophical Transactions* [**1**, 2:376–479]. In so doing, he was, in a sense, maximizing the publicity for the new result; he gave it to the most important scientific bodies in both his native England and in France. Moreover, after the founding of the London Mathematical Society in 1865, he took yet

another opportunity to highlight his work on the rule by lecturing at one of the Society's early meetings on the proof he had subsequently discovered of the general case [**1**, 2:498–513].

This mathematical triumph was followed in the late 1860s through the mid 1870s by a troubled period which found Sylvester first casting about for mathematical direction, next in premature retirement from his post at the Military Academy, and then adrift and unemployed in London. All of this changed in 1876, when he accepted the first professorship of mathematics at The Johns Hopkins University in Baltimore, Maryland. He had made a bold transatlantic move as a young man of twenty-eight; in 1876, at the age of sixty-one, he did it again. The second time around the outcome was much more positive.

Sylvester's arrival on American shores marked the beginning of a quarter-century-long process of establishing mathematics at the research level in the United States [27]. By 1877, Sylvester had regained his research footing and had reengaged in his earlier invariant-theoretic researches. In particular, he sought to vindicate the British approach to invariant theory that he and Cayley had developed by providing a British-style proof of Gordan's finiteness theorem of 1868 (on the history of this problem, see [**9**], [**10**], and [**26**]). Despite Cayley's supposed proof in 1856 that the number of irreducible covariants of a binary quintic form is infinite [**6**], Paul Gordan showed that this number is, in fact, finite for any given binary form [**13**]. Beginning in the late 1870s, Sylvester worked on and off to supply a British proof of Gordan's theorem. Writing to William Spottiswoode on 19 November, 1876, he made his nationalistic and personal motivations crystal clear in announcing what he thought was a proof of Gordan's theorem. "The piratical Germans Clebsch and Gordan who have so unscrupulously done their best to rob us English of all the credit belonging to the discoveries made in the New Algebra will now suffer it is to be hoped the due Nemesis of their misdeeds," Sylvester declared.

> Nothing in Clebsch and Gordan is *really new* but their Cumbrous method of *limiting* (not *determining*) the Invariants to any given form.... I see a splendid vista of investigations open to me on this subject destined I believe to reduce to annihilation all that the school of Clebsch and Gordan, by aid of methods borrowed by the Germans without acknowledgement from Cayley and myself, have attempted on the subject [**42**; Sylvester's emphasis].

In Sylvester's view, this was a priority issue of the greatest magnitude. Virtually his entire scientific reputation rested on his work in invariant theory, and the Germans had not only failed to give it its due but also isolated and patched a major hole in the entire British approach. May the better theory win; this battle would be fought in the international mathematical arena.

In 1877, Sylvester was convinced that he had not only won the battle but had in some sense won the invariant-theoretic war against the Germans. In May of that year, Hermite communicated the following announcement to the Paris Academy:

> Baltimore—Since my last communication, please inform the Academy that I have resolved the problem of finding the complete set of *Groundforms* for arbitrary *forms* in n variables" [**17**, p. 975, our translation].

This stunning announcement was rather quickly followed by a retraction, but Sylvester persisted in the struggle for the result. In the fall of 1878, he met his

adversary head on when he claimed again to have the theorem in its full generality. Once again, though, his claim was false.

Sylvester continued on and off to try to find a British-style proof of Gordan's theorem, and, while he never succeeded, he also never seemed to question the personal and nationalistic motivations behind such a quest, accepting fully the reality of the increasingly international arena [**29**] in which mathematicians competed. When David Hilbert gave the first inklings in 1888 and 1889 [**19**, 2:176–198] of the new invariant-theoretic methods he would bring forth in their fuller glory in 1890 [**19**, 2:199–257], Sylvester knew he had lost a major skirmish and conceded defeat gracefully in a letter to Felix Klein. Hilbert "has rendered a very good service to Algebra, in obtaining so simple a proof of Gordan's theorem," Sylvester told Klein, "and I should like to be able to congratulate him on his brilliant invention. What a relief from the previous methods of proof!" [**41**]. International competition was important, but it was also important that credit be given where due. It is not hard to imagine, however, that Sylvester got some satisfaction from the fact that, in a real sense, Gordan had been bested as well!

As a professor at The Johns Hopkins University, Sylvester did not seek to shield his students from the realities of international competition like his own with the German invariant-theoretic school. Rather, he worked to instill in them his strong sense of the importance of international, as opposed to merely national, exposure. From his own personal experience, he knew how important the foreign imprimatur was to the establishment of real reputation. In 1881, for example, when Sylvester and his Hopkins students were hard at work on what would become their groundbreaking paper on partition theory [**1**, 4:1–81], one of the students, Fabian Franklin, devised a wonderfully simple, graphical proof of Euler's pentagonal number theorem [**12**]. Sylvester was so pleased by and excited about this result that he had Franklin write it up for communication through Hermite to the Paris Academy. On 29 April, 1881, Hermite wrote to Sylvester offering his own praise for Franklin's proof and giving an indication of the attention that it was getting in France. "It certainly will not be unpleasant for you to hear that I was not the only person to be very interested in Mr. Franklin's very original and ingenious proof," he wrote.

> Mr. Halphen, one of our most eminent young mathematicians, ... found Franklin's method so remarkable that he lectured on it in one of the recent sessions of the *Société philomatique*. Please tell Mr. Franklin that his talent is appreciated, as it deserves to be, by the mathematicians of the old world [**16**, our translation].

Writing to Cayley perhaps on the very day he received Hermite's letter, Sylvester could not contain his pleasure over this reaction to his student's work. "Hermite," he gushed, "is overflowing with admiration at the beauty of the method" [**37**].

Following the success of his student, Franklin, and in the wake of the concerted combinatorial research in which he had engaged all of his students [**23**], Sylvester began to feel the strain of leading America's first research-level program in mathematics. In 1883, after seven-and-a-half years in Baltimore, he resigned from his Hopkins professorship to assume the Savilian Chair of Geometry at Oxford; the repeal of the Universities Test Act in 1871 allowed Sylvester, as a non-Anglican, to hold the Oxford chair. At the age of sixty-nine, he returned home, his efforts at reputation-building having finally yielded what for him was the ultimate prize—the recognition of the Oxbridge academic establishment.

Sylvester's career can hardly be considered typical of mathematicians in Victorian Britain. For one thing, the fact that he was Jewish initially closed many of the usual avenues of a mathematical career to him. Twice this led him to leave Britain in the hopes of greener pastures in the United States, to broaden his focus to an international arena albeit the wrong one at least in the 1840s. Moreover, it forced him to look beyond the English academic scene for ways to establish his reputation in his chosen field. If he could not have the validation of a prestigious position at an English university, he could at least measure his self-worth in terms of an international renown—hard won and carefully cultivated—that the majority of the mathematical practitioners within the English academic system would never enjoy.

Another aspect of Sylvester that made him atypical of Victorian mathematicians was his ego. It was undeniably large. He *wanted* to be known for his research; he *wanted* his work appreciated; he was often the first person to pronounce his latest theorem "remarkable" or "beautiful." Yet, Sylvester was much more complicated than this simplistic analysis would suggest. He loved mathematics and believed it to be eternal and transcendent. He wanted to make enduring contributions to it. As he explained to Lord Brougham in the aftermath of one of his altercations with the authorities at Woolwich, "I trust... to leave a lasting mark on 'The Algebra of the Future'" [**35**]. This, too, motivated him to make connections with mathematicians internationally as well as nationally. They were all striving toward the same goal, the construction of what Camille Jordan called the "temple of Algebra" [**21**] or, more generally, the temple of mathematics. And, even if they competed like Sylvester and Gordan, they labored in common cause.

Both of these atypical aspects of Sylvester as Victorian mathematician led, in his case, to the formation, maintenance, and utilization of an international network. Beginning at mid-century, national mathematical research communities—defined in terms of specialized professional associations, specialized journals, venues for the training of future researchers, and the overall emphasis on the production of original research—were under formation in Europe and somewhat later in the United States [**25**]. Sylvester, also beginning at mid-century, seemed to have a strong sense of the value of a further step in the professional development of mathematics, the internationalization of the field.

ACKNOWLEDGMENTS. The present article is part of a larger joint project on the internationalization of mathematics in the nineteenth century.

We thank Professor François Jongmans for making copies of the Sylvester-Bienaymé correspondence available to the first author and for his correspondence with both authors. We are also grateful to the following institutions for their kind hospitality and for permission to quote from their archives: The Master and Fellows of St. John's College, Cambridge; The Johns Hopkins University; the Archives of the Paris *Académie des Sciences*; and University College London Library.

REFERENCES

1. H. F. Baker, ed., *The Collected Mathematical Papers of James Joseph Sylvester*, 4 vols. Cambridge: University Press, 1904–1912; reprint ed., New York: Chelsea Publishing Company, 1973 (all page references to Sylvester's papers refer to the pagination in this edition).
2. Irenée-Jules Bienaymé to J. J. Sylvester, draft of a letter dated 4 April, 1852, in the private collection of Arnaud Bienaymé.
3. Carl W. Borchardt to J. J. Sylvester, 6 April, 1852, St. John's College, Cambridge, Sylvester Papers, Box 2.
4. Carl W. Borchardt, "Développements sur l'équation à l'aide de laquelle on détermine les inégalités séculaires du mouvement des planètes," *Journal de mathématiques pures et appliquées* **12** (1847), 50ff, or Carl W. Borchardt, *Gesammelte Werke*, ed. G. Hettner, Berlin: Georg Reimer, 1888, pp. 15–30.

5. Umberto Bottazzini, "Il diciannovesimo secolo in Italia," in Dirk Struik, *Mathematica: Un profilo storico*, Bologna: Il Mulino, 1981, pp. 249–312.
6. Arthur Cayley, "A Second Memoir Upon Quantics," *Philosophical Transactions of the Royal Society of London* **146** (1856), 101–126, or *The Collected Mathematical Papers of Arthur Cayley*, ed. Arthur Cayley and A. R. Forsyth, 14 vols., Cambridge: University Press, 1889–1898, 2:250–275.
7. Michel Chasles to J. J. Sylvester, 26 August, 1852, St. John's College, Cambridge, Sylvester Papers, Box 3.
8. Michel Chasles, *Traité de géométrie supérieure*, Paris: Bachelier 1852.
9. Tony Crilly, "The Rise of Cayley's Invariant Theory (1841–1862)," *Historia Mathematica* **13** (1986), 241–254.
10. ______, "The Decline of Cayley's Invariant Theory (1863–1895)," *Historia Mathematica* **15** (1988) 332-347.
11. Gotthold Eisenstein, "Extrait d'une Lettre adressée à M. Charles Hermite," *Journal de mathématiques pures et appliquées* **17** (1852), 473–477, or Gotthold Eisenstein, *Mathematische Werke*, 2 vols., New York: Chelsea Publishing Company, 1975, 2:771–775.
12. Fabian Franklin, "Sur le développement du produit infini $(1-x)(1-x^2)(1-x^3)(1-x^4)\cdots$," *Comptes rendus* **82** (1881), 448–450.
13. Paul Gordan, "Beweis dass jede Covariante und Invariante einer binären Form eine ganze Function mit numerische Coefficienten einer endlichen Anzahl solchen Formen ist," *Journal für die reine und angewandte Mathematik* **69** (1868), 323–354.
14. Ivor Grattan-Guinness, *Convolutions in French Mathematics*, 1800–1840: *From the Calculus and Mechanics to Mathematical Analysis and Mathematical Physics*, 3 vols., Basel: Birkhäuser Verlag, 1990.
15. D. A. Grave, "My Life and Scientific Activity," *Istoriko-Matematicheskie Issledovaniia* **34** (1993), 219–246 [in Russian].
16. Charles Hermite to J. J. Sylvester, 29 April, 1881, The Johns Hopkins University, The Milton S. Eisenhower Library, Special Collections Department, Gilman Papers, Coll #1 Corresp., Box 16.
17. Charles Hermite, "Études de M. Sylvester sur la théorie algébrique des formes," *Comptes rendus* **84** (1877), 974–975.
18. Otto Hesse, "Über die Elimination der Variabeln aus drei algebraischen Gleichungen von zweiten Grade mit zwei Variabeln," *Journal für die reine und angewandte Mathematik* **28** (1844), 68–96.
19. David Hilbert, *Gesammelte Abhandlungen*, 3 vols., Berlin: Springer-Verlag, 1932–1935.
20. François Jongmans and Eugene Seneta, "The Bienaymé Family History from Archival Materials and Background to the Turning-Point Test," *Bulletin de la Société royale des Sciences de Liège* **62** (1993), 121–145.
21. Camille Jordan to J. J. Sylvester, 13 May, 1877, St. John's College, Cambridge, Sylvester Papers, Box 2.
22. Erwin Neuenschwander, "Der Aufschwung der italienischen Mathematik zur Zeit der politischen Einigung Italiens und seine Auswirkungen auf Deutschland," *Symposia Mathematica* **27** (1986), 213–237.
23. Karen Hunger Parshall, "America's First School of Mathematical Research: James Joseph Sylvester at The Johns Hopkins University 1876–1883," *Archive for History of Exact Sciences* **38** (1988), 153–196.
24. ______, "Chemistry through Invariant Theory? James Joseph Sylvester's Mathematization of the Atomic Theory," in *Experiencing Nature: Proceedings of a Conference in Honor of Allen G. Debus*, ed. Paul Theerman and Karen Hunger Parshall, Boston/Dordrecht: Kluwer Academic Publishers, in press, pp. 81–111.
25. ______, "Mathematics in National Contexts (1875–1900): An International Overview," in *Proceedings of the International Congress of Mathematicians: Zürich*, ed. S. D. Chatterji, 2 vols., Basel/Boston/Berlin: Birkhäuser Verlag, 1995, 2:1581–1591.
26. ______, "Toward a History of Nineteenth-Century Invariant Theory," in *The History of Modern Mathematics*, ed. David E. Rowe and John McCleary, 2 vols., Boston: Academic Press, Inc., 1989, 1:157-206.
27. Karen Hunger Parshall and David E. Rowe, *The Emergence of the American Mathematical Research Community, 1876–1900: J. J. Sylvester, Felix Klein, and E. H. Moore*, Providence: American Mathematical Society and London: London Mathematical Society, 1994.
28. Adrian Rice, "Mathematics in the Metropolis: A Survey of Victorian London," *Historia Mathematica* **23** (1996).
29. Brigitte Schroeder-Gudehus, "Nationalism and Internationalism," in *Companion to the History of Modern Science*, ed. R. C. Olby *et al.*, London and New York: Routledge, 1990, pp. 909–919.

30. J. J. Sylvester to Joseph Bertrand, 12 April, 1864, Académie des Sciences, Pochette de Séance, 18 April, 1864.
31. J. J. Sylvester to Irenée-Jules Bienaymé, 7 February, 1852, in the private collection of Arnaud Bienaymé.
32. J. J. Sylvester to Irenée-Jules Bienaymé, 4 June, 1852, in the private collection of Arnaud Bienaymé.
33. J. J. Sylvester to Irenée-Jules Bienaymé, 27 August, 1852, in the private collection of Arnaud Bienaymé.
34. J. J. Sylvester to Lord Brougham, 9 August, 1854, University College London Archives, Brougham Papers, Sylvester, J. J.:20232.
35. J. J. Sylvester to Lord Brougham, 8 May, 1863, University College London Archives, Brougham Papers, Sylvester, J. J., 20253.
36. J. J. Sylvester to Arthur Cayley, undated, St. John's College, Cambridge, Sylvester Papers, Box 10 (Tony Crilly has estimated the date of this letter as between January and February, 1862).
37. J. J. Sylvester to Arthur Cayley, 12 May, 1881, St. John's College, Cambridge, Sylvester Papers, Box 11.
38. J. J. Sylvester to Pafnuty Chebyshev, 5 July, 1859 in P. L. Chebyshev, *Polnoe Sobranie Sochinenii*, vol. 5 *Prochie Sochinenia Biograficheskie Materialy*, Moscow/Leningrad: AN SSSR, 1951, p. 448.
39. J. J. Sylvester to Joseph Henry, 12 April, 1846, Smithsonian Institution Archives, Henry Papers, M099 #8573, or Nathan Reingold and Marc Rothenberg, ed., *The Papers of Joseph Henry*, 6 vols., Washington: Smithsonian Institution Press, 1972-1992, 6:407–410.
40. J. J. Sylvester to Thomas Archer Hirst, 21 March, 1863, University College London Archives, London Mathematical Society Papers, Sylvester, J. J.
41. J. J. Sylvester to Felix Klein, 24 November, 1889, Klein Nachlass XI, Niedersächsische Staats- und Universitätsbibliothek, Göttingen, or Gert Sabidussi, "Correspondence between Sylvester, Petersen, Hilbert, and Klein on Invariants and the Factorisation of Graphs 1889-1891," *Discrete Mathematics* 100 (1992):99–155 on p. 126.
42. J. J. Sylvester to William Spottiswoode, 19 November, 1876, St. John's College, Cambridge, Sylvester Papers, Box 1.

Department of Mathematics and History
University of Virginia
Kerchof Hall
Charlottesville, Virginia 22903-3199, USA
khp3k@virginia.edu

School of Mathematics and Statistics, F07
University of Sydney
Sydney, N.S.W. 2006, Australia
seneta_e@maths.su.oz.au

When she was rid of the pretense of paper and pen, phrase-making and biography, she turned her attention in a more legitimate direction, though, strangely enough, she would rather have confessed her wildest dreams of hurricane and prairie than the fact that, upstairs, alone in her room, she rose early in the morning or sat up late at night to . . . work at mathematics. No force on earth would have made her confess that. Her actions when thus engaged were furtive and secretive, like those of some nocturnal animal. . . . Perhaps the unwomanly nature of the science made her instinctively wish to conceal her love of it. But the more profound reason was that in her mind mathematics were directly opposed to literature. She would not have cared to confess how infinitely she preferred the exactitude, the star-like impersonality, of figures to the confusion, agitation, and vagueness of the finest prose.

Night and Day, by Virginia Woolf,
Harcourt, Brace, Jovanovich, pp. 45–46

The Inverse-Similarity Problem for Real Orthogonal Matrices

Geoffrey R. Goodson

INTRODUCTION. The aim of this article is to make accessible some recent results ([**1**], [**2**]) in the spectral theory of unitary operators. We do this by investigating the special case of real unitary matrices, i.e., real orthogonal matrices. Our development reinforces the importance of canonical forms and matrix decompositions in the undergraduate linear algebra curriculum.

An *orthogonal matrix* Q (with entries from any ring) is a square matrix whose transpose is its inverse ($Q^TQ = QQ^T = I$). It is known that every square matrix U with entries from any field is *similar* to its transpose, that is, $UA = AU^T$ for some nonsingular matrix A; we say that such an A is a *similarity between* U and U^T and we note that A may always be chosen to be symmetric (Theorem 1 of [**6**]; see 3.2.3 of [**4**] for the easier complex case). We are interested in the form such similarities A can have when U is real orthogonal, and what they tell us about the matrix U.

Using the orthogonal canonical form of U, we show that A may be chosen to be a real orthogonal matrix satisfying $A^2 = I$. Our main theorems (Theorems 1 and 2) give necessary and sufficient conditions for *every* real orthogonal similarity A between U and its transpose to be an involution (i.e., satisfy $A^2 = I$). Finally, we investigate the case where there exist such similarities A that are not involutions. We show (Theorem 3) that the eigenvalues of U corresponding to eigenvectors of U in the orthogonal complement of the subspace $\{x\colon A^2x = x\}$ have even multiplicity.

It may be that the results of this paper are well known, or follow easily from known results. Theorem 1, for example, follows from Theorem 2 of Taussky and Zassenhaus [**6**], which says: *Every non-singular matrix transforming U into its transpose is symmetric if and only if the minimal polynomial of U is equal to its characteristic polynomial.* A survey of results concerning the links between general matrices and their transposes is given in [**5**].

However, Theorem 3 may actually be new, and certainly the generalizations to the infinite dimensional situation, which we state in Section 2, are new (see [**1**] and [**2**]). Although the proofs of our theorems use matrix methods, their infinite dimensional analogs require the spectral theory of unitary operators.

Throughout we shall be working with real orthogonal matrices belonging to the space of all $n \times n$ complex matrices $M_{n\times n}(\mathbf{C})$. Our vectors are in $\mathscr{V} = M_{n\times 1}(\mathbf{C})$, the space of $n \times 1$ complex matrices.

1. THE SIMILARITY BETWEEN A REAL ORTHOGONAL MATRIX AND ITS TRANSPOSE. Given a real orthogonal matrix U, we can construct a real orthogonal involutory similarity A between U and its transpose in the following way:

Every real orthogonal matrix U has an *orthogonal canonical form* D (see [**4**], p.108), i.e., U can be written in the form $U = QDQ^T$, where Q is a real orthogonal

matrix and D is a block diagonal matrix of the form

$$D = \begin{pmatrix} I_{\theta_1} & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & I_{\theta_2} & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & \vdots & 0 & 0 \\ \vdots & \vdots & 0 & I_{\theta_r} & 0 & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \pm 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \pm 1 \end{pmatrix},$$

and each $I_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$ for some real θ. Since every matrix of the form I_θ is similar to its transpose via the involution $K = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, we can use such matrices as building blocks to construct a real orthogonal matrix B that gives a similarity between D and D^T. More explicitly, we write

$$D = D_1 \oplus D_2 \oplus \cdots \oplus D_k,$$

where the matrices D_i are either 2×2 matrices of the form I_θ, or 1×1 matrices of the form $[\pm 1]$. Now put

$$B = B_1 \oplus B_2 \oplus \cdots \oplus B_k,$$

where $B_i = K$ if D_i is of the form I_θ, and $B_i = [1]$ when $D_i = [\pm 1]$.

Since $B^2 = I$, we see that $A = QBQ^T$ is a real orthogonal similarity between U and U^T, and $A^2 = I$.

A real orthogonal matrix is a special type of *unitary* matrix. A complex matrix U is unitary if $UU^* = U^*U = I$, where $U^* = \overline{U}^T$, the conjugate transpose of U. The eigenvalues of a unitary matrix U are of absolute value 1. Any unitary matrix is an isometry on the vector space $\mathscr{V}$ (with the Euclidean norm), so the same is true for any real orthogonal matrix. Recall that an eigenvalue of U is *simple* if it is a non-repeated root of its *characteristic equation* $p(\lambda) = \det(U - \lambda I) = 0$. For any normal matrix, eigenvectors corresponding to different eigenvalues are orthogonal. Furthermore, if the characteristic equation has a root λ repeated r times, λ is said to be an eigenvalue of *algebraic* multiplicity r (so a simple eigenvalue has algebraic multiplicity 1). On the other hand, if for a given eigenvalue λ there are r (and no more than r) independent eigenvectors, we say that the *geometric* multiplicity of λ is r. For real orthogonal matrices (in fact, for any diagonalizable matrix, and hence for any normal, unitary, or real orthogonal matrix), the algebraic and geometric multiplicities of all eigenvalues coincide. For any real orthogonal matrix U, the characteristic polynomial $p(\lambda) = \det(U - \lambda I)$ has real coefficients, so the eigenvalues of U are real (± 1), or occur in complex conjugate pairs (necessarily of modulus 1).

We now give two lemmas. The proof of the first one is a straightforward calculation, so is omitted. The second is a special case of a classical theorem of Sylvester and is the main tool in the proof of our first theorem; for the sake of completeness we sketch its proof in a special case (see problem 9 in (2.4) of **[4]**).

Lemma 1. *Let* $K = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, *and* $K' = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. *Then B is a 2×2 real orthogonal matrix if and only if there is a real θ such that*

$$\text{(a) } B = I_\theta, \quad \text{(b) } B = KI_\theta, \quad \text{or} \quad \text{(c) } B = K'I_\theta.$$

Lemma 2. *Let F and G be $n \times n$ and $m \times m$ complex matrices, respectively, and suppose they have no eigenvalues in common. Then the matrix equation $FX - XG = C$, X an $n \times m$ matrix, has a unique solution. When $C = 0$, this solution is $X = 0$.*

Proof: We restrict our attention to the case where $C = 0$. If $FX = XG$, then $F^kX = XG^k$ for all $k = 0, 1, \ldots,$ and hence by linearity, $p(F)X = Xp(G)$ for any polynomial $p(t)$. Choose $p(t)$ to be the characteristic polynomial of F, and invoke the Cayley-Hamilton Theorem to obtain $p(F)X = 0 = Xp(G)$. Since the eigenvalues of F (which we denote by $\lambda_1, \ldots, \lambda_n$) are all different from the eigenvalues of G, $p(G) = (G - \lambda_1 I)(G - \lambda_2 I) \cdots (G - \lambda_n I)$ is non-singular and hence the equation $Xp(G) = 0$ has only the solution $X = 0$. ∎

Theorem 1. *Let U and A be real orthogonal matrices and suppose that*

$$UA = AU^T. \tag{1}$$

If all the eigenvalues of U are simple, then $A^2 = I$.

Proof: The idea is to show that equation (1) forces A to have real eigenvalues, which can only be ± 1, so $A^2 = \mathrm{I}$. To do this we use an orthogonal canonical form of U to write $U = QDQ^T$, where Q is real orthogonal and D is block diagonal:

$$D = D_1 \oplus D_2 \oplus \cdots \oplus D_k,$$

for some integer k, and every D_i, $i = 1, \ldots, k$, is either 1×1 (in which case $D_i = [\pm 1]$) or 2×2 with non-real eigenvalues (in which case it is a 2×2 real orthogonal matrix of the form I_θ, whose eigenvalues are $e^{\pm i\theta}$ with $\theta \neq m\pi$ for any $m \in \mathbb{Z}$).

Substituting $U = QDQ^T$ into equation (1) gives

$$DB = BD^T \tag{2}$$

where $B = Q^TAQ$ is a real orthogonal matrix with the same eigenvalues as A. Equation (2) gives us information about the block structure of B.

Partition $B = [B_{ij}]$ conformally with D. Then equation (2) is equivalent to the block matrix equations

$$D_i B_{ij} - B_{ij} D_j^T = 0; \qquad i, j = 1, \ldots, k. \tag{3}$$

Since all the eigenvalues of U, and hence also of D, are simple, the eigenvalues of D_i and D_j are disjoint for all $i \neq j$. By Lemma 2, $B_{ij} = 0$ for all $i \neq j$. This implies that B is a block diagonal real orthogonal matrix whose diagonal blocks are either 1×1 (in which case they are $[\pm 1]$), or 2×2 (in which case Lemma 1 ensures that they are of the form I_φ, KI_φ, or $K'I_\varphi$).

If B_{ii} is 1×1, then clearly $B_{ii}^2 = I$, so consider the case where B_{ii} is 2×2. If we are given two 2×2 real orthogonal matrices $D_i = I_\theta$ and $B_{ii} = I_\varphi$ (think of them as plane rotations), then $I_{\theta + \varphi} = I_{\varphi - \theta}$, and so $I_\theta I_\varphi = I_\varphi I_{-\theta}$ implies that $\theta + \varphi = \varphi - \theta + 2m\pi$, and hence $\theta = m\pi$. This contradiction shows that B_{ii} cannot be of the form I_φ. Consequently each B_{ii} must be of the form KI_φ or $K'I_\varphi$ (from Lemma 1). But

$$\left(KI_\varphi\right)^2 = KI_\varphi KI_\varphi = K^2 I_{-\varphi} I_\varphi = K^2 = I,$$

and the same computation shows that $(K'I_\varphi)^2 = I$. We conclude that $B^2 = I$, and hence $A^2 = I$. ∎

We now consider a partial converse to Theorem 1.

Theorem 2. *Let U be a real orthogonal matrix. If the only real orthogonal similarities between U and U^T are involutions (i.e., $UA = AU^T$ and $AA^T = I$ implies $A^2 = I$), then all the eigenvalues of U are simple.*

Proof: We show that if U has any non-simple eigenvalues, then there are real orthogonal matrices A for which $UA = A^T U$ and $A^2 \neq I$. There are essentially two cases to consider.

Case (i). Suppose that the eigenvalue $\lambda = 1$ has multiplicity at least two (a similar argument works for $\lambda = -1$). If D is an orthogonal canonical form of U, we can decompose D as $D = D_1 \oplus D_2 \oplus \cdots \oplus D_k$, where $D_j = [1]$ and $D_\ell = [1]$ for some $j \neq \ell$. Since the D_i's may be rearranged in any order, we may suppose that $\ell = j + 1 = k$, so that $D = D_1 \oplus \cdots \oplus D_{k-2} \oplus D'$, where $D' = D_{k-1} \oplus D_k$ is a 2×2 identity matrix. Now define $B = B_1 \oplus B_2 \oplus \cdots \oplus B'$, where for $i = 1, \dots, k-2$; $B_i = K$ if $D_i = I_\theta$, $B_i = [1]$ if $D_i = [\pm 1]$, and

$$B' = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

It is clear that $DB = BD^T$, but $B^2 \neq I$ since $B'^2 \neq I$. Setting $A = Q^T B Q$, we have $UA = AU^T$, but $A^2 \neq I$.

Case (ii). Suppose that U has no non-simple real eigenvalues, and that the eigenvalue $e^{i\theta}$ (with $\theta \neq n\pi$) has multiplicity at least two. As in case (i), we may assume that the last two blocks in the canonical form of D are the same, and

$$D_{k-1} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} = D_k.$$

(The other possibility is that $D_{k-1} = I_\theta = D_k^T$, and this can be treated similarly). As in case (i), construct a block matrix B, except that corresponding to $D' \equiv D_{k-1} \oplus D_k$, put $B' \equiv J$, where $J \equiv \begin{pmatrix} 0 & K \\ -K & 0 \end{pmatrix}$. Again we see that $D'J = JD'^T$ and $J^2 \neq I$, giving the desired conclusion. ■

The next result gives additional information about the nature of the eigenvalues of real orthogonal matrices. As in the proof of Theorem 1, we shall see that the eigenvalues of the matrix A play an important role.

The crucial observation is that if λ is an eigenvalue of U corresponding to some eigenvector x, then $A\bar{x}$ is also an eigenvector for U (where $\bar{x}$ is the complex conjugate of the vector x), corresponding to the same eigenvalue λ, and conversely, i.e.,

$$Ux = \lambda x \Leftrightarrow UA\bar{x} = \lambda A\bar{x}. \tag{4}$$

This is a consequence of the fact that $|\lambda| = 1$ so that $1/\lambda = \bar{\lambda}$, and if λ is an eigenvalue of any nonsingular matrix B corresponding to an eigenvector x, then $1/\lambda$ is an eigenvalue of B^{-1} corresponding to the same eigenvector x, together with the equalities:

$$UA\bar{x} = \overline{UAx} = \overline{AU^{-1}x} = \bar{\bar{\lambda}}A\bar{x} = \lambda A\bar{x},$$

The idea of the proof of Theorem 3 is to show that if x is an eigenvector of U in the orthogonal complement of the subspace $\{y \in \mathscr{V}: A^2 y = y\}$, then x and $A\bar{x}$ are orthogonal, and hence must be independent eigenvectors of U corresponding to the same eigenvalue.

Theorem 3. *Let U and A be given real orthogonal matrices. If $UA = AU^T$ and $A^2 \neq I$, then every eigenvalue of U with a corresponding eigenvector in the orthogonal complement of $H = \{x \in \mathscr{V}: A^2x = x\}$, has even multiplicity.*

Proof: We start by showing that it suffices to prove the theorem in the special cases where $U = \pm \mathrm{I}$, or $U = \tilde{I}_\theta \oplus \tilde{I}_\theta \oplus \cdots \oplus \tilde{I}_\theta$ (m copies), and $\tilde{I}_\theta = I_{\pm\theta}$ for some real number θ.

Let $D = D_1 \oplus D_2 \oplus \cdots \oplus D_k$ be an orthogonal canonical form of U, so $U = QDQ^T$. Then $DB = BD^T$, where $B \equiv Q^TAQ$.

Either $D_i = I_{\theta_i}$ for some real number θ_i, or $D_i = [\pm 1]$. We can write $D = E_1 \oplus E_2 \oplus \cdots \oplus E_m$, where E_i consists of the direct sum of all copies of those D_i having the same eigenvalues, i.e., $E_i = \tilde{I}_\theta \oplus \tilde{I}_\theta \oplus \cdots \oplus \tilde{I}_\theta$ for some real θ, or $E_i = \pm I$ (possibly different identity matrices). We do this to ensure that the spectra of E_i and E_j are disjoint for $i \neq j$; $i, j = 1, \ldots, m$.

If we partition $B = [B_{ij}]$ conformally with $D = E_1 \oplus E_2 \oplus \cdots E_m$, then

$$E_iB_{ij} = B_{ij}E_j^T; \qquad i, j = 1, \ldots, m.$$

Lemma 2 ensures that $B_{ij} = 0$ for $i \neq j$. This gives a block diagonal representation for B, and it suffices to consider each equation $E_iB_{ii} = B_{ii}E_i^T$ separately. Consequently we have essentially two cases to consider as the other cases can be treated in a similar manner.

Case (i). $U = I_\theta \oplus I_\theta \oplus \cdots \oplus I_\theta$ (m copies), for some real $\theta \neq n\pi$. Since $UA = AU^T$, both H and $H^\perp$ are U and A-invariant subspaces of $\mathscr{V}$.

We define new subspaces $\mathscr{P}_1$ and $\mathscr{P}_2$ of $\mathscr{V}$ by: $\mathscr{P}_1$ = linear span of those eigenvectors of A^2 corresponding to eigenvalues $\lambda \in S^+$, where S^+= upper half of the unit circle in the complex plane (not including ± 1). Similarly we define $\mathscr{P}_2$ in terms of S^-= lower half of the unit circle (again not including ± 1). Clearly $\mathscr{P}_1$ and $\mathscr{P}_2$ are U-invariant, for if $x \in \mathscr{P}_1$ is an eigenvector of A^2, say $A^2x = \lambda x$, then $\lambda \in S^+$, and

$$A^2Ux = UA^2x = U\lambda x = \lambda Ux,$$

so that $Ux \in \mathscr{P}_1$.

If $H_{-1} = \{x \in \mathscr{V}: A^2x = -x\}$, then it can be seen that

$$\mathscr{V} = H \oplus \mathscr{P}_1 \oplus \mathscr{P}_2 \oplus H_{-1}$$

is a direct sum of U-invariant subspaces. Furthermore, if $x \in \mathscr{P}_1$, then $A\bar{x} \in \mathscr{P}_2$ (since if $A^2x = \lambda x$, where $\lambda \in S^+$, then $A^2(A\bar{x}) = \bar{\lambda}A\bar{x}$, where $\bar{\lambda} \in S^-$).

Suppose that $\{x_1, x_2, \ldots, x_p\} \subseteq \mathscr{P}_1$ is an orthonormal set of eigenvectors of U corresponding to the eigenvalue $e^{i\theta}$. Then equation (4) implies that the set

$$\left\{A\bar{x}_1, A\bar{x}_2, \ldots, A\bar{x}_p\right\} \subseteq \mathscr{P}_2$$

is also an orthonormal set of eigenvectors corresponding to $e^{i\theta}$. We conclude that there are an even number of pairwise orthogonal eigenvectors in $\mathscr{P}_1 \oplus \mathscr{P}_2$ corresponding to $e^{i\theta}$.

In a similar manner we show that the multiplicity of $e^{i\theta}$ on the subspace H_{-1} is even. This is slightly more complicated because if $x \in H_{-1}$ is an eigenvector of U corresponding to $e^{i\theta}$, then $A\bar{x}$ is another eigenvector of U in H_{-1} corresponding to $e^{i\theta}$. However, we can use the fact that $A^2x = -x$, $x \in H_{-1}$ to show that $A\bar{x}$ and x are orthogonal.

Denote the scalar product of $x, y \in \mathscr{V}$ by (x, y) (so $(x, y) = x^*y$), and recall that the isometry A preserves this scalar product. Therefore if $x \in H_{-1}$,

$$(x, A\bar{x}) = (Ax, A^2\bar{x}) = (Ax, -\bar{x}) = -\overline{(\bar{x}, Ax)} = -(x, A\bar{x}),$$

and hence $(x, A\bar{x}) = 0$. Given an orthonormal set of eigenvectors in H_{-1} corresponding to $e^{i\theta}$, we can use this idea to construct a set of orthonormal eigenvectors of the form $\{x_1, A\bar{x}_1, x_2, A\bar{x}_2, \ldots, x_r, A\bar{x}_r\}$, that spans the same subspace. This is done by first choosing x_1, and hence $A\bar{x}_1$ as orthonormal eigenvectors. Now use the Gram–Schmidt orthonormalization process to find an eigenvector x_2 orthogonal to both x_1 and $A\bar{x}_1$ and then check that $A\bar{x}_2$ must also be orthogonal to x_1 and $A\bar{x}_1$. Continuing in this way gives a basis of eigenvectors having an even number of members.

Case (ii). If $U = I$, an argument similar to the one in case (i) gives the desired result. ■

Examples. It is interesting to notice that Theorem 3 holds even when $U = I$, the identity matrix. In fact, suppose that U and A are orthogonal matrices satisfying the conditions of Theorem 3. If they are 2×2 matrices, and if U has non-simple eigenvalues, then $U = \pm$I. If they are both 3×3 matrices, then $U = [\pm 1] \oplus [\pm 1] \oplus [\pm 1]$, (8 possibilities). It is not too difficult to list the possibilites for (the orthogonal canonical form of) U, in the 4×4 case. Here are some possibilities:

1. Let $U = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}$, so U has $\pm i$ as double eigenvalues. Consider

$$A = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix},$$

for which $UA = AU^T$, but $A^2 \neq I$.

2. Here is a general method for constructing orthogonal matrices U and A satisfying the conditions of Theorem 3, together with $A^n \neq I$ for all $n \in \mathbb{Z}$, $n \neq 0$.

First start by choosing U_0 to be any orthogonal matrix with the property that $U_0^n \neq I$ for all $n \in \mathbb{Z}$, $n \neq 0$; for example take

$$U_0 = \begin{pmatrix} \cos\pi^2 & \sin\pi^2 \\ -\sin\pi^2 & \cos\pi^2 \end{pmatrix}.$$

Now define U on $\mathscr{V} \oplus \mathscr{V}$ by

$$U = U_0 \oplus U_0^T.$$

Define the linear operator $A: \mathscr{V} \oplus \mathscr{V} \to \mathscr{V} \oplus \mathscr{V}$ by $A(x + y) = y + U_0 x$, and check that $UA = AU^T$ and $A^n \neq I$ for all $n \in \mathbb{Z}$, $n \neq 0$. From the construction it is clear that U has eigenvalues of even multiplicity.

2. INFINITE DIMENSIONAL GENERALIZATIONS. The results in this section were motivated by questions arising in ergodic theory, the study of the dynamical and spectral properties of measure–preserving transformations.

Let $U: H \to H$ be a *unitary operator* defined on a separable Hilbert space H, i.e., U is an *isomorphism* of H onto H. Measure–preserving transformations give rise to unitary operators in the following way.

We assume that $(X, \mathscr{F}, \mu)$ is a measure space that is isomorphic to the unit interval with its Borel measurable sets and Lebesgue measure. A transformation $T: X \to X$ is said to be *measure preserving* if $T^{-1}B \in \mathscr{F}$, and $\mu(T^{-1}B) = \mu(B)$ whenever $B \in \mathscr{F}$. If both T and T^{-1} are measure preserving, we call T an *automorphism* of $(X, \mathscr{F}, \mu)$. If we define an operator

$$U_T: L^2(X, \mu) \to L^2(X, \mu) \quad \text{by} \quad U_T f(x) = f(Tx); \quad x \in X; \quad f \in L^2(X, \mu),$$

then it can be shown that U_T is a unitary operator (to see it is an isometry, check first for characteristic functions of measurable sets and then extend using linearity to simple functions, etc.).

A central notion in ergodic theory is the idea of *ergodicity* of a transformation, which says, roughly speaking, that the space X is indecomposable under the action of T. More precisely, $T: X \to X$ is *ergodic* if $T^{-1}A = A$, $A \in \mathscr{F}$ implies $\mu(A) = 0$ or 1.

An important open question of ergodic theory is the nature of the spectrum of ergodic automorphisms. The first major result involving the spectra of ergodic transformations was the discrete spectrum theorem of Halmos and von Neumann ([**7**], page 68). T is said to have *discrete spectrum* if its eigenfunctions (i.e., those $f \in L^2(X, \mu)$ for which $f(Tx) = \lambda f(x)$ for some $|\lambda| = 1$) form a complete orthonormal basis of $L^2(X, \mu)$. The discrete spectrum theorem says that two ergodic transformations T_1 and T_2 with discrete spectrum are *isomorphic* (i.e., there is an automorphism S satisfying $T_1 S = S T_2$) if and only if they have the same eigenvalues. It is immediate from this theorem that if T is ergodic and has discrete spectrum, then T is isomorphic to its inverse T^{-1}.

A starting point of our work in this area was the realization that for an ergodic automorphism with discrete spectrum, every isomorphism S between T and T^{-1} is an involution. It is important to note that unitary equivalence between U_{T_1} and U_{T_2} does not in general imply that T_1 and T_2 are isomorphic. The discrete spectrum theorem says that this is true when T_1 and T_2 are ergodic with discrete spectrum.

There are few general results known concerning the spectra of ergodic automorphisms. Theorems 4 and 5 resulted from a desire to obtain more such general information, and in particular to obtain new criteria for ergodic automorphisms to have a non-simple spectrum.

In the infinite dimensional case it is possible that a unitary operator has no eigenvalues (it is said to have *continuous spectrum*). The infinite dimensional analog of simple eigenvalues is the idea of simple spectrum. A unitary $U: H \to H$ is said to have *simple spectrum* if there exists $h \in H$ with $Z(h) = H$, where $Z(h)$ is the closed linear span of the subset $\{U^n h: n \in \mathbb{Z}\}$.

The following is the infinite dimensional analog of Theorem 1 (see [**1**]).

Theorem 4. *Let $U: L^2(X, \mu) \to L^2(X, \mu)$ be a unitary operator that preserves real valued functions and has simple spectrum. If A is unitary, preserves real valued functions, and satisfies $UA = AU^{-1}$, then $A^2 = I$.*

There are very large classes of examples that satisfy the conditions of this theorem and some examples resulting from ergodic theory are given in [**1**]. Note that it can be shown that if U_T has simple spectrum, then T is ergodic. This is because 1 is always an eigenvalue for U_T, and must be a simple eigenvalue if U_T has simple spectrum, i.e., the only eigenfunctions are constants. It is now an exercise to show that ergodicity is equivalent to the fact that the only invariant functions for T (i.e., $f(Tx) = f(x)$ a.e.) are constants. We should mention that the infinite dimensional analog of Theorem 2 is also true.

The infinite dimensional analog of Theorem 3 gives additional information. This result was first formulated and proved in [**2**]. For a unitary operator $U: H \to H$ on a separable Hilbert space H, U is completely determined up to unitary equivalence by a measure σ defined on the unit circle S^1, called the *maximal spectral type* of U, and a function $\rho: S^1 \to \mathbb{Z}^+ \cup \{\infty\}$, called the *multiplicity function*. The *essential values* of this function are the values it takes almost everywhere with respect to σ (see [**3**]). We consider ∞ as an even number.

Theorem 5. *Suppose that $U, A: L^2(X, \mu) \to L^2(X, \mu)$ are unitary operators that preserve real valued functions and satisfy $UA = AU^{-1}$. Then in the orthogonal complement of the subspace*

$$\{f \in L^2(X, \mu): A^2(f) = f\},$$

the essential values of the multiplicity function of U are even.

ACKNOWLEDGMENTS. I wish to thank John Chollet for his interest in this paper and for helping me with useful suggestions and references. I am indebted to the referees and editor, whose comments have led to a considerable improvement in the paper.

I also wish to thank my son Garth who has recently completed a first course in linear algebra at Carnegie Mellon University. It was my desire to explain to him what my research was about in terms he can understand that led to this paper.

REFERENCES

1. G. R. Goodson, A. del Junco, M. Lemańczyk and D. J. Rudolph. Ergodic transformations conjugate to their inverses by involutions. *Ergodic Theory and Dynamical Systems*. 16 (1996), 97–124.
2. G. R. Goodson and M. Lemańczyk. Transformations conjugate to their inverses have even essential values. *Proc. Amer. Math. Soc.* 124 (1996), 2703–2710.
3. P. R. Halmos. *Introduction to Hilbert Space*. Chelsea, New York, 1972.
4. R. A. Horn and C. R. Johnson. *Matrix Analysis*, Cambridge University Press, New York, 1985.
5. O. Taussky. The role of symmetric matrices in the study of general matrices. *Linear Algebra and its Applications* 5 (1972), 147–154.
6. O. Taussky and H. Zassenhaus. On the similarity transformation between a matrix and its transpose. *Pacific J. Math.* 9 (1959) 893–896.
7. P. Walters. *An Introduction to Ergodic Theory*. Springer-Verlag, New York, 1982.

Department of Mathematics
Towson State University
Towson, MD 21204-7097, USA.
e7m2grg@toe.towson.edu

Rethinking Rigor in Calculus: The Role of the Mean Value Theorem

Thomas W. Tucker

1. INTRODUCTION. Mathematicians have been struggling with the theoretical foundations of the calculus ever since its inception. Bishop Berkeley's attack on Newton's "ghosts of departed quantities," Euler's claim that $1 - 1 + 1 - 1 \cdots = 1/2$, Cauchy's $\varepsilon - \delta$ definition of limit, all are part of the fascinating history of this struggle (see [7]). Calculus instructors and textbooks face the same struggle, but the tack taken, although formal, is often not sensible or honest. Instead of an admission that Newton, Leibnitz, the Bernoullis, and Euler all managed quite well without any rigorous foundations, instead of the story how a rigorous calculus took mathematicians two hundred years to get right, the Mean Value Theorem is waved, like a cross in front of a vampire, to hold the difficulties at bay. The origin of the Mean Value Theorem in the structure of the real numbers is not addressed; that is much too difficult for a standard course. Maybe it is traced back to the Extreme Value Theorem, but the trail ends there. The result is that a technical existence theorem is introduced without proof and used to prove intuitively obvious statements, such as "if your speedometer reads zero, you are not going anywhere" (if $f' = 0$ on an interval, then f is constant on that interval). That's the sort of thing that gives mathematics a bad name: assuming the nonobvious to prove the obvious. And by the way, there is nothing obvious about the Mean Value Theorem without the hypothesis of continuity of the derivative. Cauchy himself was never able to prove it in that form.

I have serious reservations about the need for formal theorems and proofs in a standard calculus course. On the other hand, for those mathematicians who do feel that need, I have a suggestion for an alternative theoretical cornerstone to replace the Mean Value Theorem (MVT); I hope textbook authors adopt it. It is much easier to state, much more intuitively obvious, and much more powerful than most mathematicians realize. It is simply this:

The Increasing Function Theorem (IFT). *If $f' \geq 0$ on an interval, then f is increasing on that interval.*

Here, *increasing* means that if $c \leq d$, then $f(c) \leq f(d)$. This would usually be called nondecreasing, but that term is awkward; for example, nondecreasing and not decreasing mean different things. It seems to make more sense to use the term *strictly increasing* for the condition that if $c < d$, then $f(c) < f(d)$. A function that is increasing, but not strictly increasing, we call *weakly increasing*.

Most of the rest of this paper is concerned with the consequences of the IFT, treating it as an axiom. I will give, however, a short independent proof of the IFT, for the sake of completeness and for readers who have probably never thought of proving the IFT directly without the MVT. Of course, the IFT follows easily from the MVT. In fact, the contrapositive of the IFT is a weak form of the MVT: if $a < b$ and $f(b) < f(a)$, there is a number c, $a \leq c \leq b$, such that $f'(c) < 0$.

It is impossible to be a pioneer in territory as well-trodden as the Mean Value Theorem. Others have championed calculus without the Mean Value Theorem (see [1], [4], [6]). The first two sections of this paper follow Lax, Burstein, and Lax [9] quite closely, although unintentionally. In fact, after searching through dozens of calculus books for the Taylor remainder proof given in this paper and finally finding it in Lax-Burstein-Lax (LBL), I felt a little uncomfortable. Maybe this paper shouldn't be published and all that is needed is an announcement "Go read LBL." Then I read Grabiner [7] and found that the Taylor remainder proof given here and in LBL is actually Lagrange's original proof. I was surprised that such a simple, direct proof could have been covered over by years of second-growth jungle.

Moreover, the idea of Lagrange's proof keeps being rediscovered for special cases like $\sin x$ or $\cos x$. For example, the *Monthly* published such an article recently [2], which then generated a subsequent Editor's Note [2] citing calculus textbooks and *Monthly* articles where the idea of [2] had already been presented. None of these references noted that the same idea works for all functions; LBL is still the only book that does that, to my knowledge. And hardly anyone seems to know the idea is really Lagrange's! Under these circumstances, it appears that some dissemination is badly needed to clear up a memory lapse of generations of mathematicians. It also appears that previous calls ([4], [6]) to downplay the Mean Value Theorem have fallen on deaf ears. Perhaps the recent debates about calculus instruction have unplugged some ears and it is time to try the call again.

2. A PROOF OF THE INCREASING FUNCTION THEOREM. There is a reasonably elementary proof of the IFT that depends only on the nested interval property of the reals: if $a_n \le a_{n+1} \le b_{n+1} \le b_n$ for all $n \ge 1$ and $\lim_{n\to\infty}(b_n - a_n) = 0$, then there is a number c such that $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = c$. The proof of the IFT given here does not require the continuity of f' and is so self-contained that it probably could be given in a standard calculus course. Although I generated this proof in response to some remarks of Peter Lax, I should have known the proof is too natural to be original. In revising this paper, I discovered Richmond's article [10], which contains essentially the same proof, and as I already knew, Ampère and Cauchy used the key observation in their own proofs.

Proof of the IFT. The proof depends on the following simple

Observation. Given a function f, define slope(a, b) to be the usual quotient $(f(b) - f(a))/(b - a)$. If slope$(a, b) = m$ and c is between a and b, then one of slope(a, c) and slope(c, b) is greater than or equal to m and one is less than or equal to m. For a proof, draw the obvious picture.

Suppose now that $f'(x) \ge 0$ on $[a, b]$ and that f is not increasing; that is, for some a_1, b_1 with $a \le a_1 < b_1 \le b$, we have $f(a_1) > f(b_1)$. Let $m =$ slope(a_1, b_1). Note that $m < 0$. By repeated bisection and our observation, we can find a nested sequence of intervals $[a_n, b_n]$ with slope$(a_n, b_n) \le m$ and $\lim_{n\to\infty}(b_n - a_n) = 0$. Let $c = \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n$ (the possibility $c = a$ or $c = b$ causes no difficulty). Since $f'(c) \ge 0$ and $m < 0$, for all x sufficiently near c, slope$(x, c) > m$. Thus for all large enough n, slope$(a_n, c) > m$ and slope$(c, b_n) > m$, which contradicts our observation and the fact that, by construction, slope$(a_n, b_n) \le m$. If $a_n = c$ or $b_n = c$, the contradiction is immediate. ■

As we have observed, the contrapositive of the IFT is an existence statement that if f is not increasing on the interval $[a, b]$, there exists a number c between a

and b where $f'(c) < 0$. The preceding proof is constructive, in that once one finds $a_1 < b_1$ with $f(a_1) > f(b_1)$, the bisection procedure effectively computes a number c such that $f'(c) < 0$.

3. IMMEDIATE CONSEQUENCES OF THE IFT. We first consider some consequences and variations of the IFT.

Theorem 1. *The following statements are consequences of the IFT. Assume f is differentiable on $[a, b]$ and $a < b$.*

a) *If $f'(x) \leq 0$ on the interval $[a, b]$, then f is decreasing on the interval $[a, b]$.*
b) *If $f'(x) = 0$ on the interval $[a, b]$, then f is constant on the interval $[a, b]$.*
c) *If $f'(x) > 0$ on the interval $[a, b]$, then f is strictly increasing on the interval $[a, b]$.*
d) *If $f'(x) \leq g'(x)$ on the interval $[a, b]$, then $f(x) - f(a) \leq g(x) - g(a)$ for all x in $[a, b]$.*
e) *If $m \leq f'(x) \leq M$ on the interval $[a, b]$, then $m(x - a) \leq f(x) - f(a) \leq M(x - a)$ for all x in $[a, b]$.*

Proof:

(a) Multiplication by -1 reverses inequalities and interchanges "increasing" and "decreasing".
(b) By the IFT and (a), it follows that f is both (weakly) increasing and (weakly) decreasing on $[a, b]$. That means f is constant.
(c) By the IFT, f is increasing. Suppose that $a \leq c < d \leq b$ and $f(c) = f(d)$. Since f is increasing on $[c, d]$ we must have $f(x) = f(c) = f(d)$ on $[c, d]$. That is, f is constant on $[c, d]$. Therefore $f'(x) = 0$ on $[c, d]$, contradicting $f'(x) > 0$ on $[a, b]$.
(d) Apply the IFT to $h(x) = g(x) - f(x)$ to conclude $g(a) - f(a) \leq g(x) - f(x)$.
(e) Apply (d) to $f(x)$ and Mx to get the right inequality and to mx and $f(x)$ to get the left inequality.

Theorem 1c could be called the Strictly Increasing Function Theorem (SIFT). Lax-Burstein-Lax [**9**] calls it the Criterion for Montonicity. There the IFT is derived directly from the SIFT by looking at $f(x) + mx = g(x)$, for all positive slopes m. If $f'(x) \geq 0$, then $g'(x) > 0$, so by the SIFT g is strictly increasing. Thus if $x > a$, then $f(a) + ma < f(x) + mx$. Since this inequality holds for all $m > 0$, it follows that $f(a) \leq f(x)$, that is, f is increasing. I feel, however, that this proof is a little tricky. Although the idea of perturbing a function is important throughout analysis, it comes out of the blue for a first-year calculus student. I prefer the IFT over the SIFT as a theoretical cornerstone. First, our proof that the IFT implies the SIFT is easier and more natural than a proof that the SIFT implies the IFT. More importantly, Theorem 1c, which could be called the Constant Function Theorem, follows immediately from the IFT; the only way the SIFT can get this fundamental result is via the IFT. By the way, I view the Constant Function Theorem as even more basic than the IFT. It would be nice to use it as our theoretical cornerstone, but I know of no way to use it to get the IFT.

Theorem 1d is called the Racetrack Principle by Jerry Uhl: if one car goes faster than another, it travels farther during any time interval. It is used as a theoretical cornerstone in the text [**5**].

Theorem 1e is perhaps the most important, especially from a historical viewpoint. If the inequalities are rewritten:

$$m \leq \frac{f(x) - f(a)}{x - a} \leq M$$

we have the Mean Value Inequality. The Mean Value Theorem follows immediately if we know that f' is continuous and that the Intermediate Value Theorem holds. That is exactly what Cauchy did [**7**]: he proved the Mean Value Inequality and assumed the continuity of f' and the Intermediate Value Theorem. His assumption of continuity should not be surprising since his proof of the Mean Value Inequality also assumes that the difference quotient $(f(x+h) - f(x))/h$ approaches $f'(x)$ uniformly as h approaches 0. Peter Lax has argued that, for the theoretical foundations of an introductory calculus course, one should always avoid pathology and assume uniform continuity and uniform convergence, just as Cauchy did. It is interesting to note that before Cauchy, Ampère [**7**] saw the importance of the Mean Value Inequality and even used it as the defining property of the derivative. One could argue in a similar vein that the Mean Value Theorem should be the defining property of the derivative; Andrew Gleason has told me that a calculus textbook by Donald Richmond around 1960 did exactly that, but I have been unable to find the book.

Finally, I should comment on the hypothesis of differentiability at the endpoints, both in the IFT and in Theorem 1. All one need assume is continuity at the endpoints, just as in the MVT. Simply observe in the proof of the IFT that the initial points a_1 and b_1 can be chosen so that $a < a_1 < b_1 < b$, since if $f(a) > f(b)$ then by continuity $f(a_1) > f(b)$ for $a_1 > a$ near enough a, and $f(a_1) > f(b_1)$ for $b_1 < b$ near enough b.

4. ERROR BOUNDS AND ERROR BEHAVIOR FOR TAYLOR POLYNOMIALS.

If Theorem 1e is rewritten

$$f(a) + m(x-a) \leq f(x) \leq f(a) + M(x-a),$$

we see a glimmering of an error bound for Taylor polynomials. The proof we are about to give is almost too transparent and simple to believe: just antidifferentiate repeatedly the inequality $f^{(n+1)}(x) \leq M$. Not only does the proof give the Lagrange form of the error bound, it also creates the Taylor polynomial itself. Moreover, as we have observed, it is Lagrange's original proof and can be found in LBL [**9**]. It is also the proof I wrote for the textbook of the Calculus Consortium Based at Harvard [**8**]. On the other hand, I have so far been unable to find it anywhere else. All the other proofs I know involve applications of Rolle's Theorem to rather elaborate auxiliary functions or repeated integration by parts or clever tricks with varying parameters. None are natural and none are likely to be discovered or appreciated by an average calculus student.

Theorem 2. (Taylor Error Bound). *Suppose that $m \leq f^{(n+1)}(x) \leq M$ on the interval $[a,b]$, where $f^{(i)}$ denotes the ith derivative of f. Then on $[a,b]$*

$$m\frac{(x-a)^{n+1}}{(n+1)!} \leq f(x) - T_n(x) \leq M\frac{(x-a)^{n+1}}{(n+1)!},$$

where $T_n(x)$ is the degree n Taylor polynomial for f centered at $x = a$.

Proof: To get the upper bound, we apply Theorem 1d (the Racetrack Principle) to $f^{(n)}(x)$ and Mx (since $f^{n+1} \leq M$), which gives

$$f^{(n)}(x) - f^{(n)}(a) \leq M(x-a).$$

Applying the Racetrack Principle again, we get

$$f^{(n-1)}(x) - f^{(n-1)}(a) - f^{(n)}(a)(x-a) \leq M\frac{(x-a)^2}{2},$$

and again

$$f^{(n-2)}(x) - f^{(n-2)}(a) - f^{(n-1)}(a)(x-a) - f^{(n)}(a)\frac{(x-a)^2}{2} \leq M\frac{(x-a)^3}{3!}.$$

Applying the Racetrack Principle a total of $n+1$ times gives the upper bound. The lower bound is obtained the same way. ■

Theorem 2 gives error bounds only for $x \geq a$. To get similar bounds for $x \leq a$, we observe that if f is increasing and $x \leq a$, then $f(x) \leq f(a)$, rather than $f(a) \leq f(x)$. Thus for $x \leq a$, each application of Theorem 1d reverses the inequalities, but since Theorem 2 sandwiches the error for $x \geq a$, reversing inequalities will simply sandwich the error again for $x \leq a$ (although which bound is the upper one depends on whether n is odd or even). The usual two-sided error bound involving absolute values then follows immediately.

It is possible for students to discover Theorem 2 for themselves. Consider the following problem. A particle is traveling along the x-axis with position $x = f(t)$ and suppose the initial position, velocity, and acceleration are all 0. If $f'''(t) \leq 5$ for $t \geq 0$, find an upper bound on the position at time $t = 2$. Since students are well-trained to antidifferentiate acceleration to get velocity and velocity to get position, it is not unnatural to see them argue as follows:

$$\begin{aligned} f'''(t) &\leq 5 \\ a = f''(t) &\leq 5t + c_1, \quad \text{and here } c_1 = 0 \text{ since } f''(0) = 0 \\ v = f'(t) &\leq 5\frac{t^2}{2} + c_2, \quad \text{and here } c_2 = 0 \text{ since } f'(0) = 0 \\ s = f(t) &\leq 5\frac{t^3}{6} + c_3, \quad \text{and here } c_3 = 0 \text{ since } f(0) = 0. \end{aligned}$$

Thus, we get $f(2) \leq 5 \cdot 2^3/6 = 20/3$. This is a legitimate argument as long as one can justify antidifferentiating inequalities in the same way as equalities. That is exactly the point of the Racetrack Principle!

Acceleration and velocity are not a bad way of introducing Taylor series. The usual formula students memorize from physics,

$$s = s_0 + v_0 t + \frac{1}{2}at^2,$$

is precisely the degree 2 Taylor polynomial for $s(t)$ when the constant acceleration a is interpreted as the acceleration at time 0. This fact seems worth exploiting, but I don't know any textbook that makes the connection.

Taylor's theorem is usually presented as a method of bounding the error in approximating a function by its degree n Taylor polynomial. This viewpoint is particularly appropriate in studying the error for fixed x as $n \to \infty$, as in the proof of the convergence for all values of x for the Taylor series for e^x or $\sin x$.

Nevertheless, I believe that this viewpoint is overemphasized and that the true power of Taylor series is in explaining error (or convergence) behavior for fixed n as $x \to a$. Why is Simpson's Rule so much better than the Trapezoid Rule? What makes the approximation $\sin x \approx x$ so good? For numerical behavior, the important thing to know is the order of convergence for fixed n under normal circumstances and what situations might affect that order of convergence. The real point of Taylor's theorem is that the error is order $n + 1$ in $(x - a)$ with a constant depending on the $(n + 1)^{\text{st}}$ derivative.

To be more precise, we say $E(h)$ is asymptotic to Ch^n, denoted $E(h) \sim Ch^n$, if $\lim_{h \to 0} E(h)/h^n = C$. Also, we say $E(h)$ is order n with bound M if $\limsup |E(h)/h^n\}| \leq M$. Then Taylor's theorem can be viewed this way:

Corollary. *Let $E(h)$ be the error $f(x) - T_n(x)$ where $T_n(x)$ is the nth degree Taylor polynomial for f at $x = a$ and where $h = x - a$. If $f^{(n+1)}$ is continuous at $x = a$, then $E(h) \sim f^{(n+1)}(a)\ h^{n+1}/(n + 1)!$. If $|f^{(n+1)}(x)| \leq M$ in a neighborhood of $x = a$, then $E(h)$ is order $n + 1$ with bound $M/(n + 1)!$.*

5. ERROR BEHAVIOR FOR NUMERICAL INTEGRATION. Another application of the Mean Value Theorem is to explain the error behavior for various common numerical integration rules: Left Rule, Right Rule, Trapezoid Rule, Midpoint Rule, Simpson's Rule. This behavior is best described using Taylor series in Δx for the error. Numerical analysis texts sometimes do this, but calculus texts don't. Since this approach is not so well-known, I'll give a version.

The idea is to concentrate on one panel of the subdivided area. Without loss of generality, we can assume the panel is centered at the origin. Thus we wish to compute

$$I(h) = \int_{-h}^{h} f(x)\, dx, \quad \text{where } h = \Delta x/2.$$

The estimate for this single panel by the left-rectangle rule is

$$I(h) \approxeq L(h) = 2h(f(-h)).$$

The other estimates are given by

$$\begin{aligned}
\text{Left: } L(h) &= 2hf(-h) \\
\text{Right: } R(h) &= 2hf(h) \\
\text{Midpoint: } M(h) &= 2hf(0) \\
\text{Trapezoid: } T(h) &= (L(h) + R(h))/2 \\
\text{Simpson: } S(h) &= (2M(h) + T(h))/3
\end{aligned}$$

The formula relating Simpson's Rule to the midpoint and trapezoidal rules is not as well known as it should be. Students can be led to guess the weighted mean as a better estimate, if they spend a little time looking at the error behavior of the midpoint and trapezoidal rules.

We want to compute the Taylor series centered at $a = 0$ for all these functions. For the rules, this is simply a matter of replacing $f(h)$ or $f(-h)$ by the Taylor series for f centered at $a = 0$. For $I(h)$, we observe that by the Fundamental Theorem of Calculus, $I'(h) = f(h) + f(-h)$. Thus $I''(h) = f'(h) - f'(-h)$, $I'''(h) = f''(h) + f''(-h)$, etc.

The Taylor series for $I(h)$ is therefore

$$I(h) = 2f(0)h + 2f''(0)\frac{h^3}{3!} + 2f''''(0)\frac{h^5}{5!} + \cdots.$$

The series for the rules are

$$L(h) = 2h\left[f(0) + f'(0)(-h) + f''(0)\frac{(-h)^2}{2!} + \cdots\right]$$

$$R(h) = 2h\left[f(0) + f'(0)h + f''(0)\frac{h^2}{2} + \cdots\right]$$

$$M(h) = 2h[f(0)]$$

$$T(h) = 2h\left[f(0) + f''(0)\frac{h^2}{2} + f''''(0)\frac{h^4}{4!} + \cdots\right]$$

$$S(h) = 2h\left[f(0) + f''(0)\frac{h^2}{6} + f''''(0)\frac{h^4}{3\cdot 4!} + \cdots\right].$$

The error behavior for each rule is obtained by subtracting the Taylor series for $I(h)$ from the Taylor series for the rule and looking for the first term that doesn't cancel. The errors behave asymptotically as follows:

$$\text{Left Error} \sim -2f'(0)h^2$$

$$\text{Right Error} \sim 2f'(0)h^2$$

$$\text{Midpoint Error} \sim -2f''(0)\frac{h^3}{3!}$$

$$\text{Trapezoid Error} \sim 2f''(0)\left(\frac{1}{2} - \frac{1}{6}\right)h^3 = 2f''(0)\frac{h^3}{3}$$

$$\text{Simpson Error} \sim 2f''''(0)\left(\frac{1}{3\cdot 4!} - \frac{1}{5!}\right)h^5 = 2f''''(0)\frac{h^5}{180}.$$

The error behavior of these rules for the entire interval is obtained by multiplying by the number n of subdivisions and replacing h by $\Delta x/2$ where $\Delta x = (b-a)/n$, except for Simpson's rule where $h = \Delta x$. We have to replace $f^{(k)}(0)$ by a bound M_k on $|f^{(k)}|$ for the entire interval. Using $2nh = (b-a)$, we find the absolute value of the errors have the following behavior in terms of Δx:

Left: order 1 with bound $(b-a)(1/2)M_1$

Right: order 1 with bound $(b-a)(1/2)M_1$

Midpoint: order 2 with bound $(b-a)(1/24)M_2$

Trapezoid: order 2 with bound $(b-a)(1/12)M_2$

Simpson: order 4 with bound $(b-a)(1/180)M_4$.

The typical textbook problem on numerical integration is to find the value of n that guarantees the error is within a specified tolerance. In practice, one simply keeps doubling n until the desired number of digits seems to have stabilized. Thus, error behavior, rather than error bounds, may be what we really are interested in.

For example, it is useful to know that increasing n by a factor of 10 for the Left or Right rule, decreases the error by a factor of $1/10$, that is, it gives one more significant digit. Thus if it takes 1 second for a graphing calculator to compute an integral accurate to 2 digits using the Left or Right Rule, it will take 10^{10} seconds to get 12 digits of accuracy (that's 3169 years and, as my students have observed, a lot of batteries). By contrast, Simpson's Rule gets 4 extra digits for 10 times the work, and the same integral can be computed to 12 digits of accuracy in a minute or two on the same calculator (Simpson's Rule probably would get a headstart of 4 or 5 digits in the first second).

The dependence of error behavior on the higher derivatives of the integrand is also important, because it is a warning to look out for integrals whose integrand has an unbounded derivative on the interval of integration. For example, even using Simpson's Rule on $\int_0^1 \sqrt{1-x^2}\,dx$ to get an approximation for $\pi/4$ is painfully slow going. Indeed, the order of convergence is $3/2$ rather than 4.

Taylor series can be used in the same way to analyze the error behavior for numerical differentiation approximations:

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}$$

$$f''(x) \approx \frac{f(x+h) + f(x-h) - 2f(x)}{h^2}.$$

For example, students are often curious why some graphing calculators use the second of these approximations as a numerical derivative rather than the more familiar first approximation. Taylor series give the answer immediately: the second error for the approximation is order 2 while the first is order 1. The dependence of the error of each approximation on higher derivatives of f also has interesting effects. Try plotting the error near $x = 0$ with $h = .01$ for the second approximation to f', when f is the innocuous-looking function $f(x) = x^{8/3}$.

6. THE FUNDAMENTAL THEOREMS OF CALCULUS. The proof given in [**8**] for the Taylor error bound appeals to the Fundamental Theorem of Calculus to turn the inequality $f^{n+1}(x) \le M$ into the inequality $f^{(n)}(x) - f^{(n)}(a) \le M(x-a)$. I suspect this is the natural inclination of most mathematicians, and it shows how much under-appreciated the IFT is. No definite integrals are needed; the IFT itself is a disguised form of integration. The subtle connection between the IFT and the Fundamental Theorem of Calculus is worth discussing.

There are of course two main versions of the Fundamental Theorem of Calculus. There are also variations on what restrictions are placed on the integrand f. I will assume f is continuous. The theorems then are

First Fundamental Theorem of Calculus (FTC I). *If f is continuous on the interval $[a, b]$ and $F(x) = \int_a^x f(t)\,dt$ for x in $[a, b]$, then $F'(x) = f(x)$.*

Second Fundamental Theorem of Calculus (FTC II). *If f is continuous and $F(x) = f(x)$, then $\int_a^b f(t)\,dt = F(b) - F(a)$.*

The First Fundamental Theorem is not directly related to the IFT. The hard part of the proof is showing that continuous functions are Riemann integrable. The

rest is a straightforward consequence of the integral version of the Mean Value Inequality:

$$m(b-a) \leq \int_a^b f(x)\,dx \leq M(b-a),$$

where $m \leq f(x) \leq M$ on the interval $[a, b]$. Note that unlike the Mean Value Inequality for derivatives, this inequality follows easily from the definition of the Riemann integral, so easily that it is not uncommon to view the inequality as a defining property of the definite integral (the corresponding view for the Mean Value Inequality for derivatives, Ampère notwithstanding, is much less common).

On the other hand, the Second Fundamental Theorem is closely connected to the IFT. The IFT for continuously differentiable functions follows directly from the FTC II and the fact that the integral of a nonnegative function is nonnegative. In fact, that is the way the IFT is proved in [**8**]. There, the FTC II, as embodied in the relation between velocity and change in position, is taken as the intuitively clear, theoretical cornerstone, and the IFT is derived from it. I suspect, however, that most students see the IFT as more "obvious" than the FTC II.

Conversely, the IFT implies the FTC II by the method used in many calculus books: simply invoke the FTC I with $x = b$ and observe that, by the IFT (the Constant Function Theorem, Theorem 1b), two antiderivatives of f differ by a constant.

The assumption of continuity in the FTC I is necessary. The assumption of continuity in the FTC II is another matter. Of course, if $F' = f$ is not continuous, the integral might not exist. For example, if $F(0) = 0$ and $F(x) = x^2 \sin(1/x^2)$ for $x \neq 0$, then F' exists everywhere but is not even Lebesgue integrable on $[0, 1]$. Suppose, however, that we assume only that $\int_a^b f(t)\,dt$ exists. Then the familiar argument using the Mean Value Theorem still works. Just represent $F(b) - F(a)$ as a telescoping sum and use the MVT on each term of the sum to turn it into a Riemann sum for $\int_a^b f(t)\,dt$. Here the IFT does not work. Just as the MVT follows from the IFT only under the assumption of continuity of the derivative, the FTC II follows from the IFT only under the assumption of continuity of the integrand.

7. CONCLUSION. Many calculus textbooks have sections where the author is writing on automatic pilot, just putting in material demanded by users. These sections have the same dreary examples; little is new, or thought over fresh from the start. This shouldn't be surprising, since writing a calculus textbook is a significant project and one can't devote the same enthusiasm and energy to all parts of the project. I have always felt that the theoretical sections of standard calculus textbooks are most prone to such a pedestrian treatment. Moreover, calculus instruction does not place much emphasis on those theoretical sections, at least when it comes to testing. For example, a study of the compendium of final exams in [**11**] reveals only one question (out of more than 300 on 23 exams) involving the Mean Value Theorem, and that one asked for the value of c satisfying the conclusion of the Mean Value Theorem for a quadratic function. When both textbooks and instruction appear to be just going through the motions with theory, it surprises me that some critics of new textbooks like [**8**] bemoan the absence of the Mean Value Theorem or a $\epsilon - \delta$ definition of limit.

I sympathize with yearnings for an occasional foray into the theoretical structure of the calculus. I just ask that it be thoughtful and sensible. Use intuitive definitions. If a theorem is to be used without proof, like the Mean Value Theorem, keep it as simple and as "obvious" as possible. Don't use tricky proofs or

deus-ex-machina auxiliary functions. Don't prove things in more generality than necessary; even analysts don't usually deal with the discontinuous derivatives allowed by the Mean Value Theorem.

In this paper, I have tried to give a sensible approach to the Mean Value Theorem and its usual applications to monotonicity, Taylor error bounds, quadrature error bounds, and the Fundamental Theorems of Calculus. One standard application of the MVT I have not considered is l'Hopital's Rule; for a non-MVT approach, see [**3**]. LBL [**9**] has some other applications to concavity and the second derivative test for extrema.

In recent years, calculus content and pedagogy have been rethought completely. People have found that there is nothing sacred about related rates and the lecture method. It is time as well to rethink the theory taught in standard calculus classes. There is nothing sacred about the Mean Value Theorem.

ACKNOWLEDGMENTS. I wish to thank Andy Gleason, Peter Lax, and Jerry Uhl for numerous suggestions and corrections for this article. In particular, the proof given for the IFT was instigated by a bisection proof Lax showed me for the SIFT. He also showed me applications to l'Hopital's Rule, the Corrected Midpoint Rule for quadrature, and the definition of volumes and arclengths using antiderivatives rather than definite integrals; all of this I hope he puts into print. I am indebted to Gleason, whose meticulous reading caught a number of egregious errors and whose comments cleared up my muddy thinking at numerous points.

REFERENCES

1. L. Bers, On avoiding the mean value theorem, *Amer. Math. Monthly* 74 (1967), 583.
2. D. Bo, A simple derivation of the Maclaurin series for sine and cosine, *Amer. Math. Monthly* 97 (1990), 836. Editor's Note in the *Monthly* 98 (1991), 364.
3. R. P. Boas, Lhospital's rule without mean value theorems, *Amer. Math. Monthly* 76 (1969), 1051–1053.
4. R. P. Boas, Who needs these mean-value theorems anyway?, *Two-Year College Math J.* 12 (1981), 178–181.
5. W. Davis, H. Porta, J. Uhl, *Calculus & Mathematica: Derivatives: Measuring Growth*, Addison-Wesley, 1994.
6. J. Dieudonne, *Foundations of Modern Analysis*, Academic Press, New York, 1960.
7. J. V. Grabiner, *The Origins of Cauchy's Rigorous Calculus*, MIT Press, Cambridge, 1981.
8. D. Hughes-Hallett, A. M. Gleason, et al., *Calculus*, John Wiley & Sons, New York, 1994.
9. P. Lax, S. Burstein, and A. Lax, *Calculus with Applications and Computing*, *Volume* 1, Springer-Verlag, New York, 1984.
10. D. E. Richmond, An elementary proof of a theorem of calculus, *Amer. Math. Monthly* 92 (1985), 589–590.
11. L. A. Steen, editor, *Calculus for a New Century: A Pump, Not a Filter*, MAA Notes 8, Mathematical Association of America, Washington, DC, 1988.

Department of Mathematics
Colgate University
Hamilton, NY 13346
ttucker@center.colgate.edu

Commentary on Rethinking Rigor in Calculus: The Role of the Mean Value Theorem

Howard Swann

Professor Tucker's article joins the current deconstructive attack on traditional content and methods of teaching of calculus that seems to be part of the mission of the militant wing of the 'Calculus Reform Movement.' Here the primary targets are current textbooks' efforts to present the foundations of calculus and the frequent use of the mean value theorem.

As the author remarks, the traditional presentation of the foundations of calculus is often poorly motivated and incomprehensible to most students. So in reforming the teaching of the calculus sequence, one should either omit the logical foundations or attempt to make them interesting and comprehensible. The author, who is one of the co-authors of the 'Harvard Calculus' text [2] where the first option is chosen and the concept of *mathematical* proof based on rigorous definitions is eliminated entirely, urges that we keep things as "intuitive ..., simple and obvious as possible." Various demonstrations are our new "proofs;" I use the quotation marks to make the distinction. The author's favored replacement for the Mean Value Theorem (MVT), the Increasing Function Theorem (IFT), finds its intuitive justification in an *automotive* ('Racetrack') argument. Such automotive arguments are a new addition to our pantheon of "proofs." An automotive "proof" of the IFT is 'if the speedometer on a motor-car always reports a number greater than or equal to zero, then the car must be moving (weakly) forward.' The IFT is to be treated as an 'axiom,' yet the essential first foundational question for calculus is 'What is it that a speedometer is supposed to report?' Intuition falters here, for nature has yet to provide us with a speedometer.

The author states, "The origin of The Mean Value Theorem in the structure of the real numbers is not addressed; that is much too difficult for a standard course." I agree that proofs of the extreme value theorem and other global results from basic principles do not belong in today's beginning calculus texts in the present educational climate. However, an informal discussion of human attempts to define 'number' is fascinating and accessible.

For example, although we currently use the 'real numbers,' today's students, brought up on Star Trek, are delighted with the realization that there still is the following problem: When we use real numbers to represent time t and position $p(t)$, we are led to the conclusion that in moving from $p(t')$ to $p(t'')$ we must disappear infinitely often, for there is NO instant of time 'next to' t' nor any position 'next to' $p(t')$. A variant of this problem bothered Zeno 2500 years ago; it has not been resolved; the reals are indeed 'full of holes.' Why shouldn't today's students again contemplate this version of the abyss confronting human attempts to comprehend infinity, particularly when Weierstrass has contrived a remarkably clever way across?

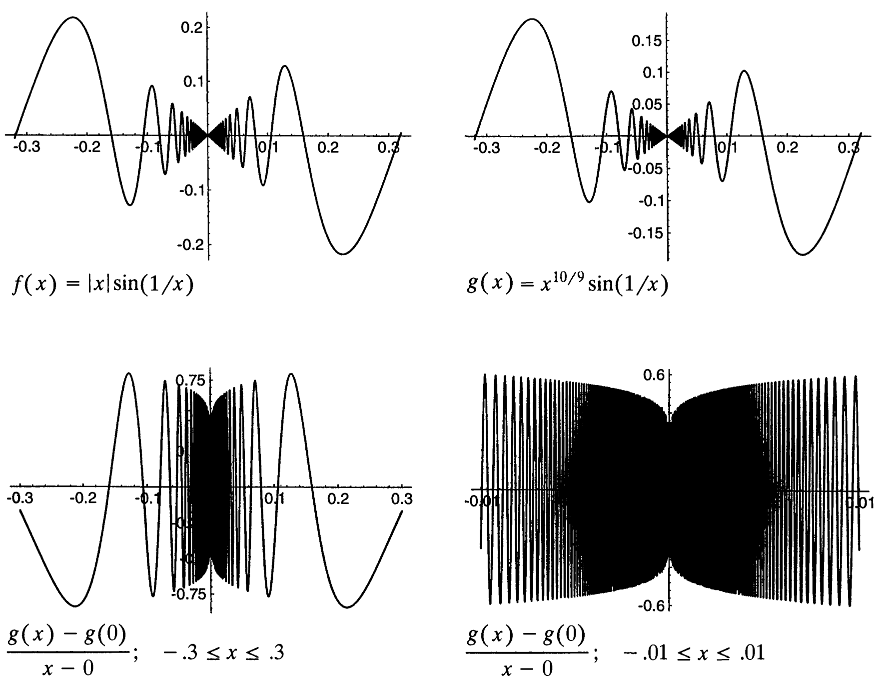
For Weierstrass, in treating continuity and differentiability, insisted that we consider only functions that are accompanied by suitable ε, δ arguments. In this

class of functions he was able to show *uniqueness* (and thus define ‘correctness’) of guesses for limits and derivatives **[1]**. These notions *are* accessible to students and give the foundations for differential calculus. When we add the global results, the implications so astonished Bertrand Russell that he pronounced [**4**, p. 64]:

> ...all goes smoothly until we reach those studies in which the notion of infinity is employed—the infinitesimal calculus and the whole of higher mathematics. The solution of the difficulties which formerly surrounded the mathematically infinite is probably the greatest achievement of which our age has to boast.

Learning to understand and appreciate proofs is a gradual process; it surely is imperative to introduce the notion of mathematical proof in beginning multisemester calculus and keep it alive even though actual proofs are few. Such an introduction is essential for later mathematics courses, and students must be made aware that the assertions of mathematics can be *proved* to be true.

The bright promise of the new technology gives us a chance to explore these ideas in a striking way. For example, using—say—Mathematica, ‘zoom’ the graphs of $f(x) \equiv |x|\sin(1/x)$ (continuously extended) and $g(x) \equiv x^{2n/(2n-1)}\sin(1/x)$ (continuously extended), search for possible ‘local linearity’ and try to decide if they are differentiable at zero. For $n > 4$, the graph of $[g(x) - g(0)]/x$ has a delightful fractal quality when you magnify the domain around zero; we give examples in Figure 1. These are not ‘important’ functions in a practical sense, but a look at such graphs encourages a sense of delight and wonder concerning the difficulties of the foundations of analysis. It does not take very much time to present and discuss these ideas.



$f(x) = |x|\sin(1/x)$

$g(x) = x^{10/9}\sin(1/x)$

$\dfrac{g(x) - g(0)}{x - 0}$; $\quad -.3 \le x \le .3$

$\dfrac{g(x) - g(0)}{x - 0}$; $\quad -.01 \le x \le .01$

Figure 1

As for the mean value theorem, the author states "And by the way, there is nothing obvious about the MVT without the hypothesis of continuity of the derivative." I believe that this is not true, for here is a pictorial "proof" of the MVT:

MEAN VALUE THEOREM. If $f(x)$ is continuous on $[a, b]$ and has a derivative on (a, b), then there is some point c, $a < c < b$, such that

$$f'(c) = \frac{f(b) - f(a)}{b - a} = \text{slope of line through } (a, f(a)) \text{ and } (b, f(b)).$$

Pictorial "proof:" Intuitively, the assumptions of the theorem mean that the graph of $f(x)$ is smooth between $(a, f(a))$ and $(b, f(b))$ and presumably has no sharp corners since f has derivatives. If the graph of $f(x)$ is not a straight line, some of the graph will be above the line through $(a, f(a))$ and $(b, f(b))$ or below this line. Suppose some of the graph is above the line. Imagine a line that is parallel to the line through $(a, f(a))$ and $(b, f(b))$ but far above the line. Move it down toward the line, keeping it parallel to the line through $(a, f(a))$ and $(b, f(b))$. Since there are no corners on the graph, when the line first hits the graph at some point $(c, f(c))$, surely it will be *tangent* to the graph at such a point. So, if our definition of the derivative as the slope of a line that is *tangent* to the graph at $(c, f(c))$ is any good, the slope of this tangent line must be $f'(c)$. *But* since the line is *parallel* to the line through $(a, f(a))$ and $(b, f(b))$, it will have the same slope as this line, i.e.

$$f'(c) = (f(b) - f(a))/(b - a).$$

A similar argument holds if some of the graph is below the line. ■

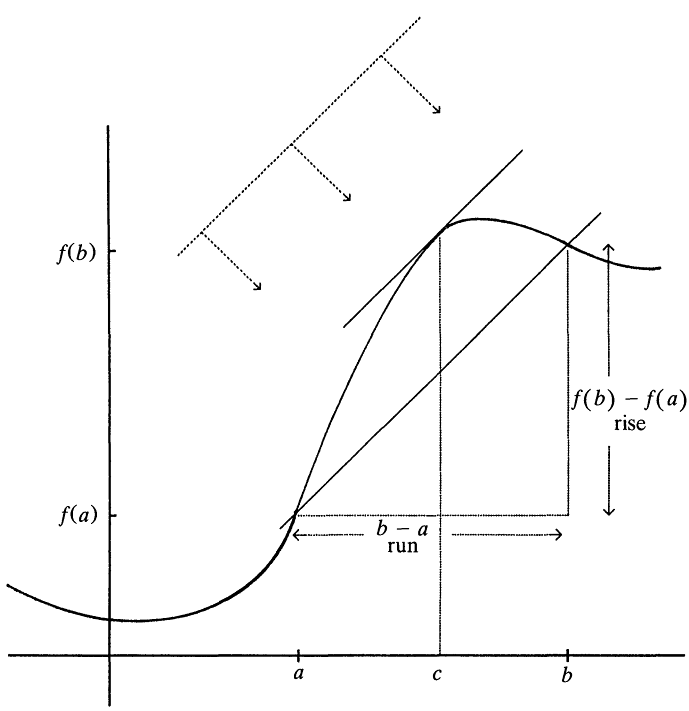


Figure 2

It is to be hoped that the phrases 'presumably' and 'if the definition is any good' make the students suspicious. This should lead to consideration of a special case (Rolle's theorem) where it is clear that the only assumption necessary is the extreme value theorem. This in turn invites a discussion (no proofs) of the extreme value theorem as one of the crucial *global* tests of the effectiveness of Weierstrass' definition of continuity. The boundedness theorem, the extreme value theorem, and the intermediate value theorem are no longer 'obvious' when students have realized that the problems with the 'holes' in the real numbers extend to any intuitive sense of continuity of a function on an interval. For example, is Weierstrass' definition of continuity strong enough to force a continuous function to be *bounded* on a closed bounded interval? The announcement that the answer is 'yes' is an excellent promotional preview of later courses. At any rate, once students have 'bought' Rolle's theorem, we can then use the conventional proof to show that the MVT must hold.

Here we reverse the author's prescription for giving mathematics a bad name; such a sequence of arguments reveals the charm and power of mathematics, for we prove that a questionable complicated result *must* be true if we assume other simpler results that are less questionable.

The author offers us a mathematical proof of the Increasing Function Theorem, "easier than most proofs of the MVT". He presents the theorem as

If $f' \geq 0$ on an interval, then f is increasing on that interval.

We infer from the proof that the interval is $[a, b]$, closed and bounded, and that we are to have one-sided derivatives at a and b.

The key observation for the author's proof is the following: Given f, let slope$(a, b) = [f(b) - f(a)]/(b - a)$. The author points out that "If slope$(a, b) = m$ and c is between a and b, then one of slope(a, c) and slope (c, b) is greater than or equal to m and one is less than or equal to m. For a proof, draw the obvious picture."

The "obvious picture" encourages this assertion, but knowing that the art of converting a "proof" to a proof is one of the key skills our majors should learn, if we are giving a proof here, we must go further. Two *mathematical* proofs are immediately discovered; a proof by contradiction (four main cases) or a direct proof. The direct proof shows first that the result must be true if $m = 0$, and then uses the same 'deus-ex-machina' auxiliary function that annoys the author when it is employed to prove the MVT from Rolle's Theorem.

The author admonishes us: "... previous calls to downplay the MVT have fallen on the deaf ears of textbook writers. Maybe calculus reform has unblocked some ears and it is time to try the call again."

I borrow a phrase from the author and wave *Occam's Razor*, that 'principle of parsimony,' "like a cross in front of a vampire, to hold the" attack on the mean value theorem at bay.

The mean value theorem is actually a friendly theorem; what has it done to provoke such ire? It provides one more test of the effectiveness of Weierstrass' definition of limit and continuity and is used endlessly to establish all sorts of results; more, by the author's admission, than the IFT. For example, how do we *prove* that the formula for arc length is correct without the MVT? We do not prove it, apparently; with the wave of the symbol '$\sim$,' we make ourselves content with the arguments of "Newton, Leibnitz, the Bernoullis and Euler."

Which is more intuitively 'obvious' and persuasive; the pictorial "proof" of the MVT that we sketched above, or an automotive "proof" of the IFT?

A one-line proof shows that the author's IFT follows from the MVT. The author's suggested proof for the MVT from the IFT requires, in addition to the usual assumptions for the MVT, that the function's derivatives be extendible to a continuous function on the closed interval, requires the extreme value theorem and the intermediate value theorem, and fails to establish that the sought-for value for c is strictly between points a and b. This is essential, for example, for showing that we can *repeat* the application of L'Hospital's rule a second time in evaluating a limit.

One positive note: The author's argument (Theorem 2) for an error bound for Taylor's series is elegant and should be adopted by one and all.

However, I do not find the main arguments of the paper to be persuasive. Those of us who, as the author says, "bemoan the absence of the Mean Value Theorem or the ε, δ definition of limit" regret that "it is time... to rethink the theory taught in standard calculus classes." Some of us are disappointed that "there is nothing sacred about related rates;" we used to regard them highly, for related rates give us the heat equation, one of the classic models of mathematical physics, and the primary example for the study of elliptic and parabolic partial differential equations.

Whether or not the militants' 'final product' is 'better,' which is by no means established [3], one thing is clear: books such as the "Harvard Calculus" are "enablers;' by legitimizing the abandonment of the concepts of mathematical proof, related rates, convergence of series, and so forth from the calculus sequence, other texts and teachers will feel free to follow.

Mathematics is unique in its concern with rigorous foundations and proofs. Here its role as 'Queen and servant of the Sciences' is to offer the content of calculus as an anchor of certainty to aid the disciplines it serves. Should we not attempt to convey some sense of the remarkable way that the results of calculus can be proved to be true to those who will use it?

REFERENCES

1. C. B. Boyer, *The History of the Calculus and its Conceptual Development*, Dover, 1959.
2. D. Hughes-Hallett, A. M. Gleason et al., *Calculus*, John Wiley & Sons, New York, 1994.
3. K. Johnson, Harvard Calculus at Oklahoma State, *Amer. Math. Monthly* 102 (1995) 794–797.
4. B. Russell, *Mysticism and Logic*, W. W. Norton & Co., Inc., New York, 1929.

Department of Mathematics and Computer Science
San José State University
San José, CA 95192
swann@sjsumcs.sjsu.edu

Fermat's Last Theorem, the Four Color Conjecture, and Bill Clinton for April Fools' Day

Edward B. Burger and Frank Morgan

1. INTRODUCTION. We often fool our students. We fool them by stating and proving theorems that are in a polished and final form. As a result, students are often unaware of the evolution of ideas, strategies, statements, proofs and even important mistakes that inevitably lead to the beautiful theorems and proofs they encounter. There is much value in understanding early ideas and mistakes. In some cases, mistakes lead to the discovery of powerful and deep new mathematical truths. Thus we believe that it is occasionally appropriate to celebrate mathematical mistakes and we cannot imagine a better time for such a celebration than on April Fools' Day. Here we describe and outline three erroneous proofs from the 19th century. We provide "proofs" of Fermat's Last Theorem, the Four Color Conjecture, and the fact that one of us is Bill Clinton. This paper is based upon a special undergraduate mathematics colloquium the authors gave on April 1, 1996.

2. FERMAT'S LAST THEOREM. The problem that was to become one of the most infamous open problems in mathematics had a most inauspicious beginning. Circa 1637, Pierre de Fermat (Figure 1a), while studying Bachet's Latin translation of *Arithmetica* by Diophantus, came upon a discussion of the Pythagorean theorem. This inspired Fermat to write the following, now famous, lines in the margin:

> "*It is impossible to separate a cube into two cubes, or a biquadrate into two biquadrates, or generally any power except a square into two powers with the same exponent. I have discovered a truly wonderous demonstration of this, which this margin is too narrow to contain.*"

Fermat was notorious for making such statements with usually little or no justification or proof. By the 1800's all of Fermat's statements had been resolved; all but the above one—his "last" one.

Fermat's Last Theorem. *For any integer $N \geq 3$, there are no integer solutions to*

$$x^N + y^N = z^N, \qquad \text{with} \qquad xyz \neq 0.$$

We do point out that Fermat himself did provide a complete proof of the above result in the case of $N = 4$. His proof involved a clever idea now known as Fermat's method of descent. The theme of 'method of descent' is to hypothesize that there are positive integer solutions to the problem at hand and then to use those solutions to construct another set of positive integer solutions that are, in some sense, smaller than the hypothesized solutions. Iterating this procedure indefinitely leads to a contradiction since there are only finitely many positive integer solutions that are smaller than the hypothesized solutions.

On March 1, 1847, Gabriel Lamé **[9]** (Figure 1b), who had already proved the result in the case $N = 7$, announced to the Paris Academy that he had a complete proof of Fermat's Last Theorem. However Joseph Liouville **[10]** quickly pointed out a serious error in the argument. The following outline of a "proof" of Fermat's Last Theorem contains some of the same mistakes that Lamé made. Can you find the errors?

Before embarking upon our "proof" of Fermat's Last Theorem, we define some notions that play a major role in the argument to follow. Let p be an odd prime and consider the polynomial $f_p(x) = x^p - 1$. If we write $\zeta_p = e^{2\pi i/p}$, then since the powers of ζ_p are zeros of f, we easily conclude that

$$f_p(x) = \left(x - \zeta_p^0\right)\left(x - \zeta_p^1\right)\left(x - \zeta_p^2\right)\cdots\left(x - \zeta_p^{p-1}\right).$$

The key to the following argument is the idea of extending the notion of integer and employing the arithmetic of generalized integers. We define the ring of p *cyclotomic integers* $\mathbb{Z}[\zeta_p]$ by

$$\mathbb{Z}[\zeta_p] = \left\{\sum_{n=0}^{p-1} a_n \zeta_p^n : a_n \in \mathbb{Z}\right\}.$$

For $\alpha, \beta \in \mathbb{Z}[\zeta_p]$, we say that α *divides* β, denoted as $\alpha | \beta$, if there exists an element $\gamma \in \mathbb{Z}[\zeta_p]$ such that $\beta = \alpha\gamma$. In this case we say that α is a *factor of* β. An element $\omega \in \mathbb{Z}[\zeta_p]$ is called a *unit* if $\omega | 1$. For example, ζ_p^t is a unit for any $t \in \mathbb{Z}$. A nonzero element $\pi \in \mathbb{Z}[\zeta_p]$ is a *prime* if π is not a unit and whenever $\pi = \omega_1\omega_2$, $\omega_1, \omega_2 \in \mathbb{Z}[\zeta_p]$, then either ω_1 or ω_2 is a unit.

"*Proof*" *of Fermat's Last Theorem.* By factoring the exponent N into primes, it is not hard to see that it suffices to prove Fermat's Last Theorem for $N = 4$ and for N an arbitrary odd prime. As we noted before, Fermat himself proved the case $N = 4$, thus we need consider only the case $N = p$, where p is an odd prime. In 1770, Euler proved the result for $p = 3$ and hence we may assume that p is a prime greater than 3.

We prove the theorem by contradiction. That is, we assume that there is an integer solution to

$$x^p + y^p = z^p, \qquad \text{with} \qquad xyz \neq 0. \tag{1}$$

By dividing out any common factors, we may assume that x, y, and z are pairwise relatively prime. We now consider the two possible cases.

Case 1. The prime p does not divide xyz.

Case 2. The prime p does divide xyz.

In order to analyze these cases, we make the following fundamental observation:

$$z^p = x^p + y^p = (-y)^p\left(-\left(\frac{x}{y}\right)^p - 1\right)$$

$$= (-y)^p f_p\left(-\frac{x}{y}\right) = (-y)^p \prod_{n=0}^{p-1}\left(-\frac{x}{y} - \zeta_p^n\right).$$

Thus,

$$\prod_{n=0}^{p-1} \left(x + \zeta_p^n y\right) = z^p. \tag{2}$$

Perhaps not surprisingly, the arithmetic of the cyclotomic integers $\mathbb{Z}[\zeta_p]$ is analogous to the arithmetic of ordinary integers $\mathbb{Z}$. For example, two elements $\alpha, \beta \in \mathbb{Z}[\zeta_p]$ are said to be *relatively prime* if they have no common factors other than units. Given this, it turns out that in **Case 1**, the factors occurring in (2) are pairwise relatively prime. In **Case 2**, however, it turns out that all the elements occurring in the product (2) have as a common factor the prime $(\zeta_p - 1)$, and once we divide out this factor from each element, the remaining cyclotomic integers are pairwise relatively prime.

Suppose our solution to (1) lies in **Case 1**. In this case, because the factors in the product of (2) are pairwise relatively prime, each element in the product must be a perfect p power of a cyclotomic integer multiplied by a unit. In particular, there must exist a nonzero cyclotomic integer ω and a unit ε so that

$$x + \zeta_p y = \varepsilon \omega^p.$$

After some work and computation, this reveals that $x \equiv y \bmod p$.

Thus, from our original equation $x^p + y^p + (-z)^p = 0$, we have discovered that $x \equiv y \bmod p$. By symmetry, our argument may be reworked to deduce $x \equiv -z \bmod p$ and $y \equiv -z \bmod p$. These congruences together with Fermat's Little Theorem, which states that $n^p \equiv n \bmod p$ for any integer n, reveal

$$0 = x^p + y^p + \left(-z\right)^p \equiv x + y + \left(-z\right) \equiv 3x \bmod p.$$

Therefore, p divides $3x$. As p is a prime greater than 3, p must divide x, but this contradicts the assumptions of **Case 1**. Hence **Case 1** is impossible.

Thus our hypothesized solution to (1) must satisfy **Case 2**. In this case we know that each of the elements in the product (2) has a common prime divisor of $(\zeta_p - 1)$ and cyclotomic integers $(x + y)(\zeta_p - 1)^{-1}$, $(x + \zeta_p y)(\zeta_p - 1)^{-1}$, $(x + \zeta_p^2 y)(\zeta_p - 1)^{-1}, \dots, (x + \zeta_p^{p-1} y)(\zeta_p - 1)^{-1}$ are pairwise relatively prime. By an analysis similar to the one in **Case 1**, we must have

$$\left(x + \zeta_p^n y\right)\left(\zeta_p - 1\right)^{-1} = \varepsilon_n \omega_n^p, \tag{3}$$

where ε_n is a unit and ω_n is a cyclotomic integer for $n = 0, 1, 2, \dots, p-1$ such that $\omega_0, \omega_2, \dots, \omega_{p-1}$ are all pairwise relatively prime. Using basic algebra and arithmetic in $\mathbb{Z}[\zeta_p]$, one may deduce the following identity free of the variables x and y:

$$\omega_1^p + \left(\tau_1 \omega_{p-1}\right)^p = \tau_2\left(\zeta_p - 1\right)^{tp} \gamma^p, \tag{4}$$

where τ_1 and τ_2 are units in $\mathbb{Z}[\zeta_p]$, $\gamma \in \mathbb{Z}[\zeta_p]$ with ω_1, ω_{p-1}, γ, $(\zeta_p - 1)$ all pairwise relatively prime, and $t \in \mathbb{Z}$ is greater than 1. Hence we have just found pairwise relatively prime cyclotomic integers X, Y, Z all relatively prime to $(\zeta_p - 1)$ such that

$$X^p + Y^p = \nu\left(\zeta_p - 1\right)^{tp} Z^p, \tag{5}$$

where ν is a unit and $t \in \mathbb{Z}$ satisfies $t \geq 1$.

We now employ Fermat's method of descent. We begin by noting that we may factor the left-hand side of (5) as before and deduce that

$$\prod_{n=0}^{p-1} \left(X + \zeta_p^n Y \right) = \nu \left(\zeta_p - 1 \right)^{tp} Z^p.$$

In this case again each factor $X + \zeta_p^n Y$ is divisible by $(\zeta_p - 1)$. We now argue as before to deduce that there exist pairwise relatively prime cyclotomic integers X_1, Y_1, Z_1, all relatively prime to $(\zeta_p - 1)$, and a unit ν_1 so that

$$X_1^p + Y_1^p = \nu_1 \left(\zeta_p - 1 \right)^{t_1 p} Z_1^p,$$

where t_1 is an integer such that $1 \le t_1 < t$. Hence we have a solution to an equation that is very similar to (5). The key difference is that in the exponent of $(\zeta_p - 1)$ we now have the integer t_1 rather than the integer t, where $1 \le t_1 < t$. We may repeat this procedure to produce pairwise relatively prime cyclotomic integers X_2, Y_2, Z_2, all relatively prime to $(\zeta_p - 1)$, a unit ν_2, and an integer t_2, $1 \le t_2 < t_1$, so that

$$X_2^p + Y_2^p = \nu_2 \left(\zeta_p - 1 \right)^{t_2 p} Z_2^p.$$

Repeating this process produces an infinite sequence of decreasing integers all between 1 and t: $1 \le \cdots < t_n < \cdots < t_2 < t_1 < t$. This absurdity implies that **Case 2** is impossible. Hence there must be no nonzero integer solutions to $x^p + y^p = z^p$ and this completes the "proof" of Fermat's Last Theorem. Where did we go wrong?

The flaw. It turns out that our sentence, "*Perhaps not surprisingly, the arithmetic of the cyclotomic integers* $\mathbb{Z}[\zeta_p]$ *is analogous to the arithmetic of ordinary integers* $\mathbb{Z}$," is inaccurate in one very important respect: in general, cyclotomic integers do not have unique factorization into primes. If you think about it, that was the key step that allowed us to conclude that each element in the product was a perfect p power; recall our statement, "*because the factors in the product of* (2) *are pairwise relatively prime, each element in the product must be a perfect p power of a cyclotomic integer multiplied by a unit.*" In fact Lamé himself wrote [**9**, p. 314]:

> Now, if one wants to make the product $k^p m m' m'' \cdots m^{(p-1)}$ equal to the pth power of a complex number C, it is necessary that the numbers $m, m', m'', \dots, m^{(p-1)}$, that do not admit a common divisor, even taken two at a time, be equal to pth powers, respectively.

Liouville then wrote [**10**, p. 319]:

> Nonetheless, some initial investigations lead me to believe that one should first try to establish, for the new complex numbers, a theorem similar to the elementary proposition for the ordinary integers, namely that there is only one way to decompose a product into prime factors.

The smallest prime for which $\mathbb{Z}[\zeta_p]$ does not satisfy the unique factorization property is $p = 23$. In particular, it is a straightforward calculation to verify that

$$\begin{aligned}\left(1 + \zeta_{23}^2 + \zeta_{23}^4 + \zeta_{23}^5 + \zeta_{23}^6 + \zeta_{23}^{10} + \zeta_{23}^{11}\right)\left(1 + \zeta_{23} + \zeta_{23}^5 + \zeta_{23}^6 + \zeta_{23}^7 + \zeta_{23}^9 + \zeta_{23}^{11}\right) \\ = 2\zeta_{23}^5\left(1 + \zeta_{23} + \zeta_{23}^2 + \zeta_{23}^4 + \zeta_{23}^5 + 3\zeta_{23}^6 + \zeta_{23}^7 + \zeta_{23}^8 + \zeta_{23}^{10} + \zeta_{23}^{11} + \zeta_{23}^{12}\right).\end{aligned}$$

It turns out that 2 is prime in $\mathbb{Z}[\zeta_{23}]$ and does not divide either of the two factors on the left-hand side of the above identity.

There is another, somewhat more technical, problem in deriving identity (4): we need to know that given any unit τ_0, there exists a unit τ_1 such that $\tau_0 = \tau_1^p$. This is not only not obvious, but sometimes impossible.

Ernst Kummer discovered that the preceding problems do not arise in certain cases. In particular, he called a prime p a *regular prime* if p does not divide the class number of $\mathbb{Z}[\zeta_p]$ (denoted by h). The class number h is a positive integer that, in a delicate sense, measures how far the ring is from having the unique factorization property ($h = 1$ if and only if we have unique factorization into primes). Thus the "proof" we have outlined is actually a *correct* proof of Fermat's Last Theorem for regular primes; this result is due to Kummer **[8]** (see **[3]**, **[11]** for further details and a complete proof). Notice that if $\mathbb{Z}[\zeta_p]$ satisfies the unique factorization property, then $h = 1$ and plainly p does not divide h. Thus, if $\mathbb{Z}[\zeta_p]$ has the unique factorization property, then Fermat's Last Theorem is true for the prime exponent p.

Kummer's contributions in the direction of Fermat's Last Theorem had tremendous ramifications. In May of 1847, Kummer **[7]** wrote a letter to Liouville and stated:

> Concerning the elementary proposition for these complex numbers, that a composite complex number may be decomposed into prime factors in only one way, which you so correctly cite as lacking in this proof—a proof defective in other ways as well—I can assure you that it does not hold in general for complex numbers of the form
>
> $$a_0 + a_1\zeta_p + a_2\zeta_p^2 + \cdots + a_{p-1}\zeta_p^{p-1},$$
>
> but it is possible to rescue it, by introducing a new kind of complex number, which I have called an ideal complex number.

Kummer's observation led to the birth of what we now call *ideals*. Finally in 1994, Andrew Wiles (Figure 1c) **[12]**, using powerful machinery from abstract algebra and the theory of elliptic curves, accomplished the momentous feat of producing a complete and correct proof for all primes.

Figure 1. (a) P. de Fermat; (b) G. Lamé; (c) A. Wiles.

3. THE FOUR COLOR CONJECTURE. In 1852, Francis Guthrie asked whether every planar map (as in Figure 2) with connected countries can be colored with at most four colors, so that adjacent countries have different colors. It was finally proved by Kenneth Appel and Wolfgang Haken [**1**] of the University of Illinois in 1976 with over a thousand hours of computer time, checking over ten thousand cases (Figure 3a).

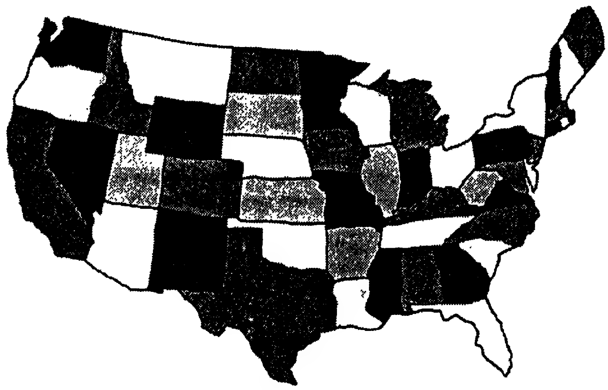


Figure 2. The Four Color Conjecture says that a planar map with connected countries can be colored with four colors such that bordering countries have different colors.

But the first published "proof" appeared in 1879, by Alfred Kempe [**6**], a London barrister and amateur mathematician (Figures 3b, 3c). The result stood for eleven years, until Percy Heawood [5] caught the error in 1890. Heawood confessed his paper's aim was "*rather destructive than constructive, for it will be shown that there is a defect in the now apparently recognized proof.*"

Figure 3. (a) W. Haken (and K. Appel) finally proved the Four Color Conjecture in 1976 with over a thousand hours of computer time (photo courtesy of the American Mathematical Society); (b) Alfred Kempe gave the first published "proof" of the Four Color Conjecture in 1879; (c) Plate from Kempe's 1879 proof of the Four Color Conjecture [**6**].

Kempe's Proof of the Four Color Conjecture. We proceed by induction on the number of countries. We remark that if the result holds for countries that meet only in threes at a point, then it holds in general, since for example if countries meet in fours at a point, they can be perturbed as in Figure 4 to meet only in threes, then colored, then restored. (Countries diagonally across from each other are allowed to be the same color.) So we may assume if we like that countries meet only in threes.

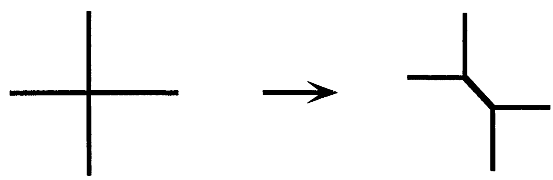


Figure 4. Countries meeting in fours can be perturbed to meet in threes.

The conjecture obviously holds for one country. We will suppose the conjecture holds for $n - 1$ countries and prove it for n countries. It is not hard to show (using the concept of Euler characteristic, for example), that some country, say "Massachusetts," has at most five neighbors. Shrink away Massachusetts as in Figure 5, and color the rest by induction. Now restore Massachusetts and find a way to color it by the following cases.

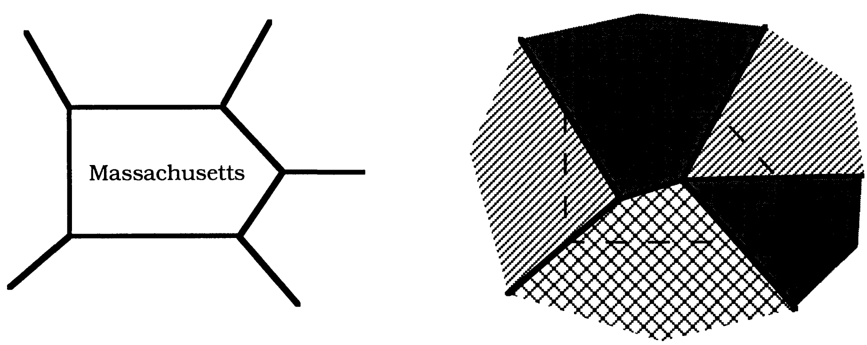


Figure 5. Shrink away the *n*th country and color the rest by induction.

Case 1. Massachusetts has fewer than four neighbors. Simply color Massachusetts with an unused color.

Case 2. Massachusetts has exactly four neighbors. We may assume they are four different colors, say green, red, blue, and yellow as in Figure 6, or we could color Massachusetts with an unused color. If there is no red-yellow chain from the top to the bottom, starting at the top, switch the colors of all connected reds and yellows. Then color Massachusetts red. If there is a red-yellow chain from top to bottom,

then there cannot be a not blue-green chain from left to right. Starting at the left, switch all connected blues and greens. Then color Massachusetts blue.

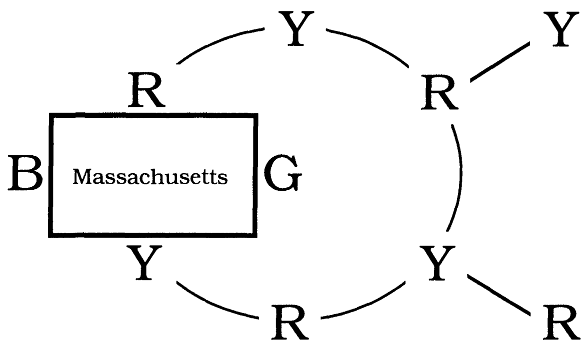


Figure 6. If there is a red-yellow chain, starting from the left switch all connected blues and greens and then color Massachusetts blue.

Case 3. Massachusetts has exactly five neighbors. Again we may assume that the neighbors use all four colors, as in Figure 7. If there is no green-yellow chain from left to right, starting from the right, switch all connected yellows and greens and then color Massachusetts yellow. If there is no green-blue chain from left to right, starting from the right, switch all connected blues and greens and then color Massachusetts blue. If there are both green-yellow and green-blue chains, starting at the top switch all connected reds and blues and starting at the bottom switch all connected reds and yellows. Then color Massachusetts red.

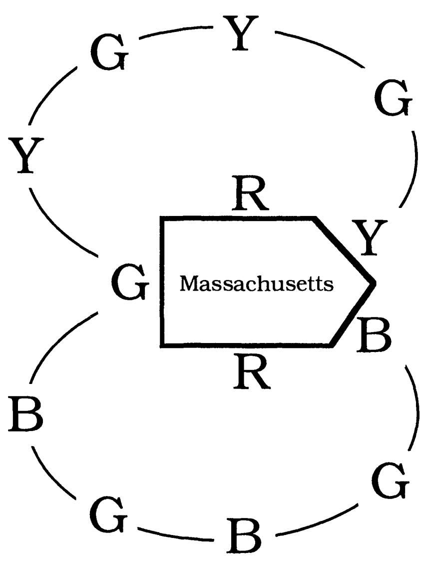


Figure 7. If there are both green-yellow and green-blue chains, starting at the top switch all connected reds and blues and starting at the bottom switch all connected reds and yellows. Then color Massachusetts red.

That completes Kempe's "proof." Before reading on, can you find the flaw?

The flaw. The problem is that the two chains of **Case 3** might cross as in Figure 8, so that the regions they enclose can overlap. Then switching reds and blues from the top may interfere with switching reds and yellows from the bottom. You might think you could still salvage the proof, but it is just not that easy. For a nice account of the Four Color Conjecture, see [**13**].

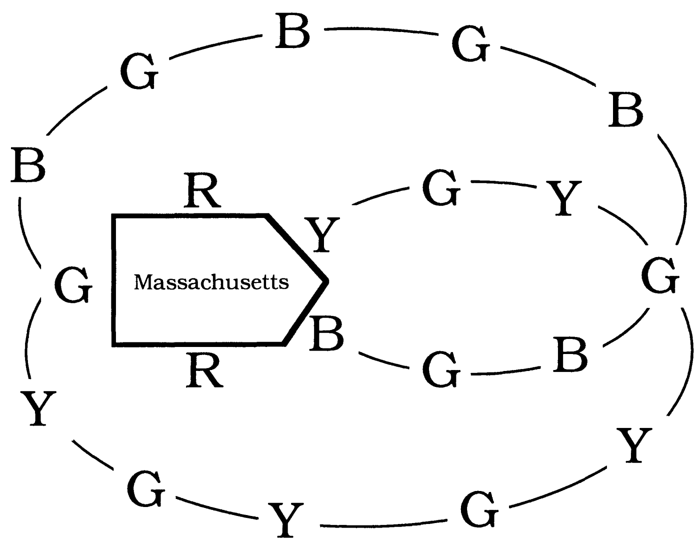


Figure 8. The unjustified assumption in Figure 7 is that the chains do not cross, but they might as illustrated here.

Incidentally, the Four Color Conjecture technically does not apply to the continental United States, even if you include water to connect a state such as Michigan, because Kentucky has a little disconnected piece surrounded by Tennessee and Missouri. If you do not believe us, just try to drive from Louisville all the way southwest to the tiny New Madrid Bend region without leaving Kentucky.

4. I AM BILL CLINTON. We now illustrate how one can "prove" anything by proving that "I am Bill Clinton." Consider the statement:

This statement is false or I am Bill Clinton.

If the statement is false, then the first clause is true, so the statement is true, which is a contradiction. Therefore the statement must be true. Since the first clause is false, the second clause must be true. Hence I am Bill Clinton. (See Figure 9.)

Actually 19th Century mathematics permitted one to make such arguments. Mathematics was restructured so as to forbid such self-referential statements, founded on a much more technical Zermelo-Frankel set theory, which does not allow sets to contain themselves; see [**4**, Section 3.4] or [**2**].

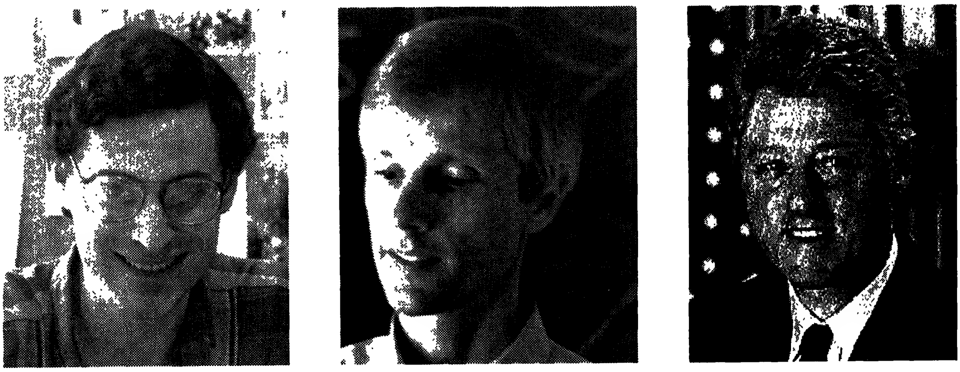


Figure 9. Authors Edward Burger and Frank Morgan and an unidentified third claimant to "I am Bill Clinton."

REFERENCES

1. K. Appel and W. Haken, Every planar map is four colorable, *Bull. Amer. Math. Soc.* **82** (1976), 711–712.
2. J. Barwise and J. Etchemendy, *The Liar*, Oxford University Press, New York, 1987.
3. H. M. Edwards, *Fermat's Last Theorem A Genetic Introduction to Algebraic Number Theory*, Springer-Verlag, New York, Heidelberg, Berlin, 1977.
4. W. S. Hatcher, *Foundations of Mathematics*, W. B. Saunders Co., Philadelphia, 1968.
5. P. J. Heawood, Map-colour theorems, *Quart. J. Math. Oxford*, **24** (1890), 322–338.
6. A. B. Kempe, On the geographical problem of the four colours, *Amer. J. Math.* **2** (1879), 193–200.
7. E. E. Kummer, Extrait d'une lettre de M. Kummer à M. Liouville, *Jour. de Math.* **12** (1847), 136.
8. E. E. Kummer, Allgemeiner Beweis des Fermat'schen Satzes, dass die Gleichung $x^\lambda + y^\lambda = z^\lambda$ durch ganze Zahlen unlösbar ist, für alle diejenigen Potenz-Exponenten λ, welche ungerade Primzahlen sind und in den Zählern der ersten $(\lambda - 3)/2$ Bernoulli'schen Zahlen als Factoren nicht vorkommen, *Jour. für Math. (Crelle)* **40** (1850), 130–138.
9. G. Lamé, Démonstration générale du théorème de Fermat, sur l'impossibilité, en numbres, de l'équation $x^n + y^n = z^n$, *C.R. Acad. Sci. Paris* **24** (1847), 310–315.
10. J. Liouville, Observations, *C.R. Acad. Sci. Paris* **24** (1847), 315–316.
11. P. Ribenboim, *13 Lectures on Fermat's Last Theorem*, Springer-Verlag, New York, Heidelberg, Berlin, 1979.
12. A. Wiles, Modular elliptic curves and Fermat's Last Theorem, *Ann. of Math.* **141** (1995), 443–551.
13. R. J. Wilson and J. J. Watkins, *Graphs: an Introductory Approach*, John Wiley & Sons, Inc., New York, 1990.

Department of Mathematics
Williams College
Williamstown, MA 01267
Edward.B.Burger@williams.edu
Frank.Morgan@williams.edu

NOTES

Edited by **Jimmie D. Lawson**

Euler's ϕ Function on Arithmetic Progressions

D. J. Newman

We will study the ϕ function's behavior on arithmetic progressions. The numerical evidence is extremely misleading in many cases. Thus it was pointed out by Dov Jarden in his book *Recurring Sequences*, Riveon Lematematika, Jerusalem, 1973, page 65, that $\phi(30n+1) > \phi(30n)$ for all $n \leq 10000$. (Indeed, further computations have shown that this persists up to 20,000,000.) Another case that resists is that of $6n+1$ vs. $6n+2$, where again, the inequality $\phi(6n+1) > \phi(6n+2)$ holds well into the millions.

At least, in this case, we were able to explicitly produce the smallest n predicted by our theorem: it is 6,197,024. In Jarden's case and in many others, the n is not explicitly available and it may be beyond the reach of any possible computers! At any rate we can prove the following.

Theorem. *If a, b, c, d are nonnegative integers with $a, c > 0$ and $ad - bc \neq 0$ then there exists a positive integer n for which $\phi(an+b) < \phi(cn+d)$.*

Remarks: By replacing b by $b + aN$ and d by $d + cN$, N large, we see in fact that our theorem gives infinitely many such n. The condition $ad - bc \neq 0$ is certainly a necessary one since otherwise we would have the case of $a = c$, $b = d$ which gives equality of $\phi(an+b)$ and $\phi(cn+d)$, but even worse we would have the case of $\phi(4n)$ which is always strictly bigger than $\phi(n)$.

Proof of the theorem: Begin with the linear Diophantine equation $ax + b = Py$, where P is chosen prime to a and with $(\phi(P)/P) < \epsilon$, ϵ later to be specified. For example, P may be chosen as the product of many consecutive primes. Notice that then

$$\phi(ax+b) = \phi(Py) \leq \phi(P)y = \frac{\phi(P)}{P}(ax+b) < \epsilon(ax+b).$$

Next observe that a general solution to our equation is given by $x = x_0 + kP$, $y = y_0 + ka$, k an arbitrary integer. This gives $cx + d = cx_0 + d + kcP$, and if we denote $\delta = \gcd(cx_0 + d, cP)$ and note that it divides $a(cx_0 + d) - y_0 cP = a(cx_0 + d) - c(ax_0 + b) = ad - bc$, we obtain $\delta \leq |ad - bc|$. Factoring out δ gives $cx + d = \delta(A + kB)$ where A and B are relatively prime.

We now recall Dirichlet's great theorem on primes in arithmetic progressions. This states that if A and B are relatively prime integers with $B > 0$, then there are infinitely many integers, k, for which $A + Bk$ is a prime. We may apply Dirichlet's theorem to our case and thereby choose k so as to make $A + Bk$ a prime, which we shall call p.

Now we have

$$\phi(cx + d) = \phi(\delta p) \geq \phi(p) = p - 1 \geq p/2 = \frac{cx + d}{2\delta}.$$

Since $(cx + d)/(ax + b)$ is monotonic and positive, it has a positive minimum, m, so

$$\frac{cx + d}{2\delta} \geq \frac{m(ax + b)}{2\delta} \geq \frac{m(ax + b)}{2|ad - bc|} = \epsilon(ax + b),$$

where ϵ is defined by this last equation. ∎

1830 Rittenhouse Square, Philadelphia, PA 19103

A Note on Weyl's Inequality

Steve Fisk

We present a simple inequality for the eigenvalues of Hermitian matrices that implies Weyl's inequality and the monotonicity theorem. The underlying idea, to intersect suitably chosen subspaces to obtain eigenvalue inequalities, is not new. See [**1**] and [**2**].

Lemma 1. *If $S_1, \dots, S_k$ are subspaces of an n-dimensional vector space V, and if* $\dim(S_1) + \cdots + \dim(S_k) > n(k - 1)$, *then the intersection of all the S_i's is non-zero.*

Proof: Consider the map from the direct sum of the subspaces S_i to the sum of $k - 1$ copies of V that sends $(v_1, \dots, v_k)$ to $(v_1 - v_2, v_2 - v_3, \dots, v_k - v_{k-1})$. The intersection of all the S_i's is the kernel of this map, and has dimension at least $\dim(S_1) + \cdots + \dim(S_k) - (k - 1)n$. ∎

If H is an n-by-n Hermitian matrix, we denote its ordered eigenvalues by $\lambda_1(H) \leq \cdots \leq \lambda_n(H)$. An n-by-n matrix X is *negative semidefinite* if $v^*Xv \leq 0$ for every vector v. For instance, the zero matrix is negative semidefinite.

Theorem 1. *Suppose $H_1, \dots, H_k$ are n-by-n Hermitian matrices such that $H_1 + H_2 + \cdots + H_k$ is negative semidefinite. Then*

$$\lambda_{i_1}(H_1) + \lambda_{i_2}(H_2) + \cdots + \lambda_{i_k}(H_k) \leq 0$$

for all $i_1, \dots, i_k \in \{1, \dots, n\}$ such that $i_1 + \cdots + i_k < n + k$.

Proof: Let S_j be the subspace spanned by eigenvectors of H_j corresponding to the eigenvalues $\lambda_{i_j}(H_j), \lambda_{i_j+1}(H_j), \dots, \lambda_n(H_j)$. Since

$$\sum_{j=1}^{k} \dim(S_j) = \sum_{j=1}^{k} (n - i_j + 1) = nk + k - (i_1 + \cdots + i_k) > n(k - 1),$$

Lemma 1 ensures that there is a unit vector x in the intersection of all the S_j's.

Now $\lambda_{i_j}(H_j)$ is the smallest eigenvalue of H_j restricted to S_j, and therefore

$$\lambda_{i_j}(H_j) \le x^* H_j x, \quad \text{for } j = 1, \dots, k$$

since each S_j is invariant under H_j. Adding these inequalities gives

$$\sum_{i=1}^{k} \lambda_{i_j}(H_j) \le x^* \left(\sum_{j=1}^{k} H_j \right) x \le 0. \qquad \blacksquare$$

Corollary 1 (Weyl's Inequality). *If A, B are n-by-n Hermitian, then*

$$\lambda_j(A) + \lambda_k(B) \le \lambda_{j+k+1}(A + B).$$

Proof: Take $H_1 = A$, $H_2 = B$, $H_3 = -(A + B)$, $i_1 = j$, $i_2 = k$, $i_3 = n - j - k$. The result follows from Theorem 1 and the fact that $\lambda_s(-C) = -\lambda_{n+1-s}(C)$ for any Hermitian matrix C. $\blacksquare$

Corollary 2 (Monotonicity Theorem). *If A, B are n-by-n Hermitian and B is positive semidefinite, then $\lambda_i(A) \le \lambda_i(A + B)$ for all $i = 1, \dots, n$.*

Proof: Take $H_1 = A$, $H_2 = -A - B$, $i_1 = i$, $i_2 = n - i + 1$. $\blacksquare$

The choice of eigenvalues in Theorem 1 is the best possible.

Theorem 2. *If $i_1 + \cdots + i_k \ge n + k$, then there are n-by-n Hermitian matrices $H_1, \dots, H_k$ such that $H_1 + \cdots + H_k$ is negative semidefinite and*

$$\lambda_{i_1}(H_1) + \cdots + \lambda_{i_k}(H_k) > 0.$$

Proof: We may assume that $i_1 + \cdots + i_k = n + k$. Let H_s be the diagonal matrix whose diagonal is all 1's, except for the entries in rows $(i_1 + \cdots + i_{s-1}) + 2 - s, \dots, (i_1 + \cdots + i_s) - s$, where the entries are $1 - k$. $H_1 + \cdots + H_k = 0$ is negative semidefinite, and $\lambda_{i_s}(H_s) = 1$, so $\lambda_{i_1}(H_1) + \cdots + \lambda_{i_k}(H_k) = k$. $\blacksquare$

REFERENCES

1. Yasuhiko Ikebe, Toshiyuki Inagaki, and Sadaaki Miyamoto, The monotonicity theorem, Cauchy's interlace theorem, and the Courant-Fisher theorem, *Amer. Math. Monthly* **94** (1987), 352–354.
2. R. C. Thompson and L. J. Freede, On the eigenvalues of sums of Hermitian matrices, *Linear Algebra Appl.* **4** (1971), 369–376.

Bowdoin College, Brunswick, ME 04100
fisk@bowdoin.edu

Minimal Polynomials Over Cyclotomic Fields

Ming-chang Kang

1. Introduction. Throughout this note, n and m denote two positive integers with d the greatest common divisor of n and m. We define e by $n = de$, and let $\zeta_n = e^{2\pi\sqrt{-1}/n}$, which is a primitive nth root of 1.

The minimal polynomial of ζ_n over $\mathbb{Q}$ is $\Phi_n(X)$, the nth cyclotomic polynomial, defined by $\Phi_n(X) := \prod_{\text{g.c.d. } \{i,n\}=1}(X - \zeta_n^i)$, by [J, p. 264]. However, if $m \geq 3$, $\Phi_n(X)$ might decompose over $\mathbb{Q}(\zeta_m)$, the field generated by ζ_m over the rational numbers. In [A, p. 531], one finds the exercise: Find the minimal polynomial over $\mathbb{Q}(\zeta_3)$ of $\zeta_6, \zeta_9, \zeta_{12}$ respectively. The answer is: g.c.d. $\{\Phi_6(X), X^2 - \zeta_3\}$, g.c.d. $\{\Phi_9(X), X^3 - \zeta_3\}$, g.c.d. $\{\Phi_{12}(X), X^4 - \zeta_3\}$ for $\zeta_6, \zeta_9, \zeta_{12}$ respectively. (Hint. ζ_6 is a zero of both $\Phi_6(X) = X^2 - X + 1$ and $X^2 - \zeta_3$. Hence it is a zero of their difference, which is $X - 1 - \zeta_3$.) One might wonder whether the analogous result is true in general. We show that this is so.

Theorem 1. *The minimal polynomial of ζ_n over $\mathbb{Q}(\zeta_m)$ is the greatest common divisor of $\Phi_n(X)$ and $X^e - \zeta_d$.*

2. Preliminaries. The Euler φ-function $\varphi(n)$ is the number of positive integers $\leq n$ that are relatively prime to n. If we denote by $(\mathbf{Z}/n\mathbf{Z})^\times$ the group of units in $\mathbf{Z}/n\mathbf{Z}$, the ring of integers modulo n, it is not difficult to see that an integer k is relatively prime to n if and only if k is invertible in $\mathbf{Z}/n\mathbf{Z}$. In particular, $\varphi(n)$ is the cardinal number of $(\mathbf{Z}/n\mathbf{Z})^\times$.

Lemma 2 [J, p. 107]. *If n_1 and n_2 are relatively prime, then $\varphi(n_1 n_2) = \varphi(n_1)\varphi(n_2)$. Moreover, if $n = p_1^{\lambda_1} \cdots p_r^{\lambda_r}$, where $\lambda_i \geq 1$ and $p_1, \cdots, p_r$ are distinct prime numbers, then $\varphi(n) = n\prod_{i=1}^r (1 - p_i^{-1})$.*

Lemma 3 (Chinese Remainder Theorem) [J, p. 107]. *If n_1 and n_2 are relatively prime, then $\mathbf{Z}/n_1 n_2\mathbf{Z} \simeq \mathbf{Z}/n_1\mathbf{Z} \times \mathbf{Z}/n_2\mathbf{Z}$. It follows that $(\mathbf{Z}/n_1 n_2\mathbf{Z})^\times \simeq (\mathbf{Z}/n_1\mathbf{Z})^\times \times (\mathbf{Z}/n_2\mathbf{Z})^\times$.*

Lemma 4. *Let d and d' be positive integers such that d is divisible by every prime factor of d'. Then (i) $\varphi(d'd) = d'\varphi(d)$; and (ii) for any non-zero integer i, g.c.d. $\{i, d\} = 1$ if and only if g.c.d. $\{i, d'd\} = 1$.*

Proof: (i) Let $d = p_1^{\lambda_1} \cdots p_r^{\lambda_r}$ be the decomposition of d into the product of prime numbers with $\lambda_i \geq 1$, $p_i \neq p_j$ if $i \neq j$. Then $d'd = p_1^{\varepsilon_1} \cdots p_r^{\varepsilon_r}$ for some ε_i where $\varepsilon_i \geq \lambda_i$ for $1 \leq i \leq r$. By Lemma 2, we find $\varphi(d'd) = d'd\prod_{i=1}^r(1 - 1/p_i)) = d'\varphi(d)$.

(ii) This is easy since a prime number p divides d if and only if it divides $d'd$. ■

3. A proof of Theorem 1. Let l be the least common multiple of n and m, $n = de$, and $d =$ g.c.d. $\{n, m\}$. Write $e = d'e'$ and $m = df$, where g.c.d. $\{d, e'\} = 1$ and d is divisible by every prime factor of d'. Note that g.c.d. $\{e, f\} = 1$. Hence g.c.d. $\{e', m\} = 1$.

Lemma 5. $\mathbb{Q}(\zeta_n, \zeta_m) = \mathbb{Q}(\zeta_l)$.

Proof: The multiplicative subgroup generated by ζ_n and ζ_m is contained in the cyclic group generated by ζ_l. On the other hand, since $d =$ g.c.d $\{n, m\}$, we can find integers a and b such that $an + bm = d$. Dividing by nm, we get $a/m + b/n = 1/l$. Hence ζ_l belongs to the subgroup generated by ζ_m and ζ_n. ■

Lemma 6. $[\mathbb{Q}(\zeta_n, \zeta_m) : \mathbb{Q}(\zeta_m)] = [\mathbb{Q}(\zeta_n) : \mathbb{Q}(\zeta_d)] = d'\varphi(e')$.

Proof: $[\mathbb{Q}(\zeta_n) : \mathbb{Q}(\zeta_d)] = [\mathbb{Q}(\zeta_n) : \mathbb{Q}]/[\mathbb{Q}(\zeta_d) : \mathbb{Q}] = \varphi(dd'e')/\varphi(d) = \varphi(e')\varphi(d'd)/\varphi(d) = d'\varphi(e')$ by Lemma 4 and noting that e' and $d'd$ are relatively prime.

On the other hand, Lemma 5 ensures that it is sufficient to show that $[\mathbb{Q}(\zeta_l): \mathbb{Q}(\zeta_m)] = d'\varphi(e')$. Now, since $l = me$, we have $\varphi(l)/\varphi(m) = \varphi(me)/\varphi(m) = \varphi(md'e')/\varphi(m) = \varphi(md')\varphi(e')/\varphi(m) = d'\varphi(e')$ again by Lemma 4 and noting that md' and e' are relatively prime. ∎

Now to the proof of Theorem 1.

In general, the minimal polynomial of ζ_n over $\mathbb{Q}(\zeta_m)$ is a factor of the minimal polynomial of ζ_n over $\mathbb{Q}(\zeta_d)$. However, we know that the degrees of these two polynomials are the same because of Lemma 6. Thus, we can consider only the minimal polynomial of ζ_n over $\mathbb{Q}(\zeta_d)$ in the sequel. Since this polynomial divides the g.c.d. of both

$$\Phi_n(X) = \prod_{\text{g.c.d.}\,\{i,\,n\}=1} (X - \zeta_n^i) \quad \text{and} \quad X^e - \zeta_d = \prod_{0 \le j \le e-1} (X - \zeta_n^{1+jd}),$$

and the degree of this minimal polynomial is $d'\varphi(e')$, proving Theorem 1 is equivalent to showing that the set $S := \{1 + jd : 0 \le j \le e - 1,\ \text{g.c.d.}\ \{n, 1 + jd\} = 1\}$ consists of $d'\varphi(e')$ elements.

Write S as

$$S = \bigcup_{0 \le t \le d'-1} \{1 + (te' + k)d : 0 \le k \le e' - 1, \text{g.c.d.}\,\{1 + (te' + k)d, n\} = 1\}.$$

By Lemma 3, we have $(\mathbf{Z}/n\mathbf{Z})^\times \simeq (\mathbf{Z}/d'd\mathbf{Z})^\times \times (\mathbf{Z}/e'\mathbf{Z})^\times$. Thus $1 + (te' + k)d$ is invertible in $\mathbf{Z}/n\mathbf{Z}$ if and only if it is invertible in both $\mathbf{Z}/d'd\mathbf{Z}$ and $\mathbf{Z}/e'\mathbf{Z}$. Since $1 + (te' + k)d$ is invertible in $\mathbf{Z}/d\mathbf{Z}$, it is automatically invertible in $\mathbf{Z}/d'd\mathbf{Z}$ by part (ii) of Lemma 4. Thus g.c.d. $\{1 + (te' + k)d, n\} = 1$ if and only if g.c.d. $\{1 + kd, e'\} = 1$, and therefore S can be rewritten as $S = \bigcup_{0 \le t \le d'-1}(te'd + T)$, where $T := \{1 + kd : 0 \le k \le e' - 1,\ \text{g.c.d.}\ \{1 + kd, e'\} = 1\}$.

From the definition of T, we may interpret the elements in T as those invertible elements in $\mathbf{Z}/e'\mathbf{Z}$ that are of the form $1 + kd$ for some k because g.c.d. $\{d, e'\} = 1$ and we find that $\mathbf{Z}/e'\mathbf{Z} = \{1 + kd : 0 \le k \le e' - 1\} (\text{mod } e')$. Note that for any invertible element s in $\mathbf{Z}/e'\mathbf{Z}$, there is a unique integer $k (\text{mod } e')$ such that $s = 1 + kd\ (\text{mod } e')$ because we may solve the equation $dX = s - 1\ (\text{mod } e')$ uniquely. It follows that $T = (\mathbf{Z}/e'\mathbf{Z})^\times$ under such an interpretation. Hence T has precisely $\varphi(e')$ elements, and therefore S consists of $d'\varphi(e')$ elements. ∎

Finally, an exercise for the reader: Use Theorem 1 to find the minimal polynomial of ζ_n over the maximal real subfield of $\mathbb{Q}(\zeta_m)$, i.e., the field $\mathbb{Q}(\cos(2\pi/m))$.

REFERENCES

[A] M. Artin, *Algebra*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1991.

[J] N. Jacobson, *Basic Algebra I*, W. H. Freeman, San Francisco, 1974.

Department of Mathematics
National Taiwan University
Taipei, Taiwan, Rep. of China
kang@math.ntu.edu.tw

THE EVOLUTION OF...

Edited by **Abe Shenitzer**
Mathematics, York University, North York, Ontario M3J 1P3, Canada

The Significance of Mathematics: The Mathematicians' Share in the General Human Condition

Notes for a talk given on December 5, 1978 at Fordham University

Wilhelm Magnus

1. INTRODUCTION AND HISTORICAL REMARKS. The topic I propose to discuss is a concern of the mathematician. However, it is not a topic of mathematics but of philosophy, at least if one agrees that philosophy is not itself a specialized science but a discipline that deals with the interaction of all human endeavors. Although I am a working mathematician with not more than an amateur's knowledge of philosophy, I nevertheless hope to be able to make at least some valid observations that will contribute to a better understanding of a rather complex situation.

Mathematics begins with an understanding of the abstract concept of a natural number (i.e., of the numbers 1, 2, 3,—ad inf.) and the ability to count indefinitely. Today, this understanding is practically universal and, in this sense, we may say that every human being is a mathematician. It is a curious fact that the mathematical component in the emergence of civilization is hardly ever mentioned by modern historians. I have found a reference to it only in Jacob Burckhardt. It was different in antiquity. In one of his plays, Aeschylus mentions "*ἀριθμόν, ἔξοχον σοφισμάτων*" (Number, outstanding (concept) among the ingenious inventions). And Aristotle says that everything was created by God with the exception of the concept of number, which is man's invention.

Mentioning the name of Aristotle could be the starting point for a survey of the role of mathematics and of mathematical concepts as an object of philosophical investigations, including a history of epistemology. Being a mathematician and not a philosopher, I neither can nor will discuss these things. However, I should like to touch at least briefly on the work of three eminent philosophers who assigned to mathematics an extraordinary role in their systems. They are Plato, Leibnitz, and Spinoza.

Plato considers knowledge of mathematics to be a prerequisite of citizenship. Specifically, he states that anyone who calls himself a civilized person should know that there exist incommensurable quantities in geometry. For example, it is impossible to find a unit of length such that both the side and the diagonal of a square are integral (= whole) multiples of this unit. This is indeed a surprising fact. It requires a sophisticated proof and it is something beyond the range of intuitive perception. But why should everybody know it? Plato wanted everybody

to know that some facts, even surprising ones, are absolute certainties. To understand this need for certainty, one should read the plays of Aristophanes, which exhibit the emergence of nihilism in Plato's time. If we use Nietzsche's definition of nihilism as the doctrine: "Nothing is true. Everything is permitted", we find it fully illustrated in "The Clouds". In another play, "The Birds", we see the human race entering into an alliance with the birds in order to destroy the power of the gods. It should be remembered that these plays were performed in honor of a particular god, Dionysus. In still another play, "The Frogs", this very god receives a thrashing.

Plato tried to fight nihilism by exhibiting mathematics as a source of absolute truth and certainty. Today, we know that the truths and certainties of mathematics are relative rather than absolute, so we are not in a good position to fight nihilism in this way.

Leibnitz was both a philosopher and an eminent mathematician. No one ever thought more highly of mathematics than he. According to Leibnitz, mathematics is the science that tells us what is possible. As far as the physical world is concerned, i.e., that aspect of the world that Descartes called "res extensa", this statement contains at least some truth. But Leibnitz goes further. According to him, God, the supreme mathematician, created our particular world by choosing of all possible worlds the one with the greatest plenitude and variety. In this sense, ours is "the best of all possible worlds."

The success of the exact sciences (which are based on the use of mathematics) has increased the range of our knowledge of the universe to a degree enormously beyond that available to Leibnitz. Paradoxically, this has made many of us (including myself) more modest, because our more extensive knowledge has made us more aware of the range of our ignorance. We are more reluctant than Leibnitz to make definite statements about the universe, and we certainly would not make a statement like that of Descartes who said: "Give me matter and motion, and I shall make the world once more."

Like Leibnitz, Descartes was both a philosopher and an eminent mathematician. But mathematics does not play an explicit role in his philosophy although he is extremely important for the history of the exact sciences through his dichotomy of the world into "res extensa" and "res cogitans." We still are influenced by his perception of the world. But at least one philosopher made a heroic attempt to overcome this dualism, and thought he could do so by using not mathematics proper, but at least the methods of mathematics.

Spinoza proposed to derive definite philosophical truths from self-evident statements "more geometrico", i.e., in the manner of Euclid. Although there can be no doubt that Spinoza provided deep and important insights, this is not due to his method, which does not qualify as a mathematical argument. I shall illustrate with an example taken from the first page of his main work, the "Ethics". There we find the statement "By God, I understand Being absolutely infinite." What is "absolutely" infinite? Spinoza did not know of the discovery of Georg Cantor (published in 1895) according to which there are smaller and larger infinitudes. For instance, there are more points in a finite interval on a straight line than there are natural numbers $1, 2, 3\ldots$ (ad inf). What is worse, there is an infinite sequence of infinitudes, each larger than the previous one. And assuming that there exists a largest infinitude containing all the previous ones would lead to a contradiction. Now we may be able to *live* with a contradiction, but we cannot *tolerate* it in a mathematical argument.

The lesson from this is simple enough. Before we start relying on mathematics we have to understand both its potential and its limitations.

2. WHAT IS MATHEMATICS? There exists a book by Courant and Robbins with this title. It tries to answer the question by giving examples. I shall try to give a general description first and then illustrate it with a few examples. Please note: The following remarks must not be taken for an attempt at giving an epistemological definition of mathematics. Their purpose is merely to provide an intuitive understanding of the nature of mathematics.

Mathematics deals with concepts subject to the rules of logic, in particular to the postulate of the excluded middle. There exists at least one set of concepts of this type, namely that of the natural numbers.

Comments. It is not true that all statements involve concepts that are subject to logic. If we have a green piece of cloth with a bluish tint we may be in doubt whether we should call the color green or blue-green, and we may even disagree about the name we wish to give to the color. Similarly, we cannot say that a person is either tall or not tall. Even if we give an artificial definition of tallness (say, 6 feet or more) we may run into trouble because no measurement is absolutely precise. (There is a good reason why we have something like 250000 laws in this country. The law uses strangely defined concepts and has to be more and more casuistic to make them fit reality.)

As far as I know, Nietzsche was the first to point out this fact. He claimed that only man-made concepts are subject to logic. With due respect for Aristotle I would take sides against him and Nietzsche and say with Kronecker (a 19th-century mathematician) that "God made the whole numbers. All the rest is the work of man." I hope that these remarks will suffice. I am not prepared to make statements about the "reality" of the natural numbers in philosophical (ontological) terms.

Mathematical research has two important and, I believe, unique characteristics: It involves an element of the infinite—being the only secular human activity to do so—and it produces an increasing wealth of problems with increasing abstraction.

Comments and examples. The element of the infinite in mathematics can be used to prove—in this case truly "more geometrico"—that the human mind is superior to any conceivable electronic computer. I cannot describe the arguments needed here without becoming rather technical. They are forever linked with the name of one of the greatest mathematicians of our time, Kurt Gödel (1906–1978). In an age where scientists as well as philosophers try to tell us that we are really nothing particular—a survival mechanism for our genes or, at best, a freakish and rather unpleasant animal that, after all, is just capable of doing some things a little better than the more pleasant chimpanzee—our mathematical abilities provide perhaps the simplest and strongest non-metaphysical argument for our special position in nature.

To illustrate these remarks I shall use two examples. The first one is a theorem of number theory, which can be stated as follows:

Every natural number N is the sum of the squares of at most four natural numbers. Unless $N + 1$ is divisible by 8, at most three squares suffice.

Obviously, no amount of direct calculations can prove this theorem because it involves an infinitude of numbers. The proof is neither easy nor obvious and was

given (for the first part of the theorem) in the late eighteenth century by Lagrange. It clearly illustrates what I mean by "an element of the infinite." I shall need this theorem again later when discussing the motivation of the mathematician. But for now I need another example to illustrate my remarks about abstraction, and for this purpose I shall start with the *Koenigsberg Bridge Problem*:

In 1735, the great Swiss mathematician, Euler, came across a peculiar problem and described it as follows: "In the town of Koenigsberg there is an island called Kneiphof, with two branches of the river Pregel flowing around it. There are seven bridges crossing the two branches. The question is whether a person can plan a walk in such a way that he will cross each of these bridges once but not more than once.... On the basis of the above I formulated the following very general problem for myself: Given any configuration of the river and the branches into which it may divide, as well as any number of bridges, to determine whether or not it is possible to cross each bridge exactly once."

Figure 1 shows the layout of the seven bridges of Koenigsberg. Can one stroll across each of these bridges once but no more than once? If so, how?

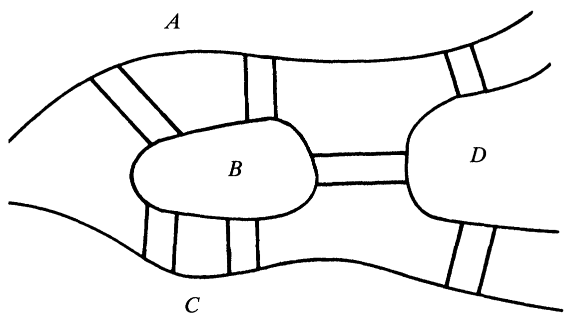


Figure 1

We begin to study this problem by throwing away unnecessary information. Since the shape and size of the islands and the countryside on the banks of the river do not matter at all, we contract each of the four areas labeled respectively A, B, C, and D to a single point. The width and shape of the bridges do not matter, either, so we replace each of them by a segment of a line or by a curve. Figure 2 shows the result of this process. This looks rather similar to a part of the subway maps exhibited in the trains in New York City, and we could rephrase the problem accordingly with stations and subway rides between them. However, we

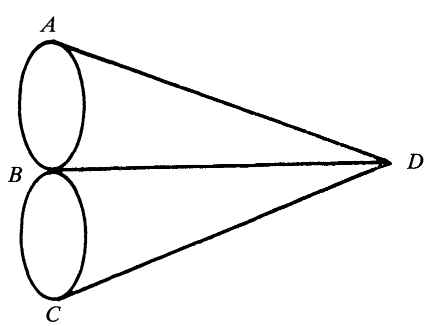


Figure 2

prefer to use the standard mathematical terminology. We shall call the whole figure a graph. The points A, B, C, D, are called *vertices*, and the connecting lines are called *edges*. Furthermore, we shall call the number of edges going through a vertex the *order* of the vertex. Obviously, the orders of A, B, C, D, are, respectively, 3, 5, 3, 3. Our problem now is: Can we find a path (i.e., a succession of edges, each having exactly one point in common with the previous one) such that this path goes through all the edges exactly once? Such a path is called an *Euler path*.

Any graph for which this question can be answered in the affirmative will be called an *Euler graph*. (This is an ad hoc notation and not common usage). Now we can show easily that our graph is *not* an Euler graph. The reason for this is the following theorem:

If a connected graph is an Euler graph then there are either no vertices of odd order or exactly two vertices of odd order.

To prove this, let us imagine that we erase any part of the path through which we have traveled in our attempt to travel through all edges exactly once. Whenever we enter a point and then leave it again, we will have to erase *two* edges through this point, reducing thereby its order by two. Therefore, if a point has odd order, then we may start our path at it, possibly pass through it several times, but then we cannot come back to it in the end, because its order would have to be even for this purpose. Therefore, if there is a point of odd order, we may try to start our path at it, but then we must end our journey at another point of odd order. (What is true for the starting point is also true for the terminal point, since we can reverse our journey). This proves our theorem, because all points other than the starting point and the terminal point must be of even order.

This is not a deep or difficult theorem, but Euler asked immediately the questions that every mathematician would ask in this situation: *Is the converse of our theorem true*? In other words: Suppose we have a connected graph in which all points are of even order. Can we start an Euler path anywhere and get back, in the end, to our starting point? And suppose we have a connected graph with exactly two points of odd order? Can we start an Euler path at one of the points of odd order and terminate it at the other one?

So far, we have established a first level of abstraction. It has led from a "finite" or "concrete" problem (the original Bridge Problem) to a general problem, and to a general theorem that covers an infinitude of "concrete" problems, namely, all possible graphs that we can draw. Now comes a second level of abstraction. We do not have to *draw* anything to define a graph. All we need is an "incidence table", which lists the points and number of edges joining any two of them. The incidence table for Figure 2 would then look as follows: The numbers in the first row simply mean that A is connected with no edge to itself, with two edges to B, with no edge to C and with one edge to D. The numbers in the other rows are similarly defined.

I hope that you will agree that incidence tables are more abstract than the graphs. But now abstraction raises immediately a new problem. We can design incidence tables without giving a graph.

Question: when does an incidence table define a graph that can be drawn in the plane so that no two edges cross each other? This is not always the case. The problem was solved in the 20th century by Kuratowski.

The reason why I brought up these things is this: In general, abstraction reduces the number of statements we can make and the number of questions we can ask. We can say more about one species of birds than about birds in general and more about birds than about the animal kingdom, etc. That the situation is different, at

least in many cases, for mathematics is a curious and, as far as I know, unique occurrence.

3. WHAT CAN MATHEMATICS DO? The functions of mathematics may be described as an extension of some of the functions of language. Before I start explaining this I have to demolish a statement that I have learned or read frequently as a sort of platitude, namely: "Mathematics *is* a language." Now even platitudes can be true, but this one happens to be sheer nonsense. It *is* true that mathematics *uses* a special language, and the reason is that our everyday language uses concepts that are not subject to logic and therefore is not suitable for the formulation of mathematical arguments and results. But a loom is not a piece of cloth, and technical terms are not ideas.

In Genesis 2, God gives the privilege of *naming* all living beings to Adam, who represents the human race. To name things—and even non-material ones like feelings and sensations—is an act of abstraction. It has been pointed out by Hans Jonas that it is also an act of image-making, another important and specifically human privilege, a sort of secondary creativity. (The ability to make images has been used by Jonas for the purpose of determining man's "specific difference" in the animal kingdom.) The ability to name things is the basis of our ability to plan and provides us with a tremendous increase of our ability to communicate, which is essential for the emergence of a coherent human society.

Now I want to show that mathematics, too, has the function of image making and that this function gives us the ability to predict. I have to be rather sketchy here, in order to avoid technicalities as far as possible.

Surprisingly, mathematics provides us with abstract images of things that are not accessible to the direct perception of our senses. At the same time, the image can be made so precise or faithful that it allows us to know all aspects of the original that are of importance to us. However, I can give here only an example, which still is not really abstract. I hope it will give at least an idea of what all of this is about.

Let me start with the fact that we can draw a map of the globe on a flat piece of paper in such a way that the map can be used for navigation. This is an important achievement of mathematics. There are many ways to do this. One of the first was found by Gerhard Kremer, who is known under the latinized form of his name Gerhardus Mercator (1512–1594). The problem is a difficult one since it is impossible to make such a map without distorting distances.

This is still a rather elementary example of the image-making power of mathematics. A much more sophisticated and much more important example is the mathematical image of an atom with a nucleus and electrons. This mathematical image consists entirely of formulas. But these formulas permit us, at least under certain circumstances, to make predictions about the behavior of the atom. This is an enormous achievement, and it is but one example of the role of mathematics in physics, chemistry, and the branches of technology based on these sciences.

Here I have to put in a word of warning. We are always tempted to overestimate the power at our disposal. In the case of language, or, specifically, of "naming", the overestimation of our power appeared in the form of incantations or conjuring. (Mathematical objects, in particular the pentagram, have been abused for the same purpose.) In the case of mathematics, it appears in the more subtle form of applying mathematical deductions to situations where this is not justified. In both cases there is involved an element of cheating, trying to get something too easily. However, I cannot try here to describe the misuses of mathematics. It is a difficult topic, and it requires careful study. But I shall discuss briefly another aspect of the

use of mathematics in physics and other exact sciences that was nicely formulated by Eugene Wigner as the title of a talk some years ago: "The unreasonable effectiveness of mathematics in the exact sciences."

Indeed, it has been a cause of astonishment for a long time that mathematics can be used at all to understand and even control the physical world. I believe that this astonishment is somewhat misplaced and that it is a consequence of Descartes' philosophy, which divides the world into "res extensa" and "res cogitans", and I should like to counter it with an aphorism by Lichtenberg (an 18th century physicist and essayist) who said: "I have found people who were astonished that cats have holes in their furs *exactly* at the places where the eyes are."

Obviously, the cat would not exist if it were otherwise. Similarly, we need order for our existence. It is therefore not so surprising that the world contains an element of order and that we have an organ to deal with it. Of course, this does not explain the extent of our ability to apply mathematics, nor does it explain the fact that the human race developed mathematics long before it became useful. So there is indeed a reason for some astonishment, which, however, should include other phenomena, such as our ability to appreciate and to create beauty and to do many other things that, at least originally, provided no visible help in the preservation of our species. I shall have to say more about these things in the next section. For now, I should like to mention a negative service that mathematics can perform. Mathematics can tell us that there are things we cannot do with the means at our disposal. For example, suppose we wish to seat the representatives, one for each, of the ca. 150 members of the United Nations at a conference table. We can not list the possible seating arrangements since their number would be greater than the number of electrons and protons in the known Universe. Of course, we are not particularly interested in such seating arrangements. But we might be interested in the arrangements of genetic material in chromosomes. I have no exact data at hand to discuss this problem, but the numbers are large, too.

4. THE PHENOMENON OF MATHEMATICS. We shall use the term "mathematics" in its strict sense: *The systematic derivation of theorems with the help of explicitly formulated arguments*. Some mathematical insights are intuitively clear, e.g., that a diameter divides a circle into two equal parts; Thales (ca. 624–548 B.C.) is supposed to have *proved* this. The fact that the side and diagonal of a square are incommensurable is not at all intuitively clear. Its discovery is ascribed to the Pythagorean school. A well-formulated proof of this and of related theorems appeared at the time of Plato and was due to his friend Theaetetos. Although Babylonian, Indian, and Chinese scholars developed a large body of mathematical knowledge, it is safe to say that mathematics in the strict sense is a creation of the Greeks. This does not mean the Athenians. With the exception of Theaetetos, none of the great Greek mathematicians lived in Athens. Euclid lived in Alexandria in Egypt. So did Apollonius. Archimedes lived in Syracuse in Sicily. Nothing like the systematic works of Euclid and Apollonius is known from other civilizations of the same or of earlier times.

What motivated these mathematicians? Not technology, not even astronomy, which, after all, was in its more sophisticated aspects not a "practical" matter at all. It is true that Archimedes developed technological applications of mathematics, in particular an instrument to compute the position of the planets. But the Romans who certainly needed and used high level technology never contributed anything to mathematics. In fact, the systematic use of mathematics for the development of technology (excluding astronomy) starts only in the 18th century.

The case for the development of mathematics was not usefulness. Earlier, we compared some functions of mathematics to some functions of language. The analogy goes even further. Language, too, is not merely an instrument of power or of usefulness. Nor is poetry. As far as mathematics is concerned, a good summary of its role appeared in an editorial (by Chandler and Edwards) in *The Mathematical Intelligencer*:

> It is a perennial problem for mathematicians to explain to the public at large what makes mathematics worthwhile if not its practicality. It is like explaining to someone who has never heard music what a lovely melody is . . . . Do let us try to teach the general public more of the sort of mathematics that they can use in everyday life, but let us not allow them to think—and certainly, let us not slip into thinking—that this is an essential quality of mathematics.
>
> There is a great cultural tradition to be preserved and enhanced. Each generation must learn the tradition anew. Let us take care not to educate a generation that will be deaf to the melodies that are the substance of our great mathematical culture.

In the past, some poets understood the beauty of mathematics. I already mentioned Aeschylus. Calderon speaks of "sublime mathematics" and Schiller calls it "divine". There are certainly more examples of this type, but they seem to become rare if not extinct in modern times. The reason for this is, of course, increasing inaccessibility of mathematics. Our latest products are available only to very few people. But the columns by Martin Gardner and occasional essays on mathematics in the *Scientific American* show that a much larger part of the population understands what mathematics is about. Some people are fascinated by Lagrange's theorem (mentioned earlier), but certainly not everybody is. However, little would be left of human civilization if we restrict it only to things that enjoy universal appreciation.

There is one more aspect of mathematics that, although well known, usually is mentioned as a mere curiosity. I believe it is more than that since it relates to the important idea of *evolution.* What are its uses? Is it a matter of pure chance or is it a response to a need, a change of conditions and environment or both? Consider the development of the exact sciences and its result, the human ability to dominate nature. In several cases, scientists found the mathematical tools they needed ready-made and available, sometimes formulated centuries earlier. The philosopher Whitehead mentions the conics, which had been thoroughly investigated by Apollonius in the third century B.C. and were available to Kepler in the 17th century A.D. The most surprising example I know of is the theory of probability. First of all, it is strange that even a situation of complete disorder, that of random events, should be subject to mathematical laws. Secondly, what provoked the study of probability was an almost universally despised human activity, namely, gambling. But one of the main contributors to the theory of probability was Pascal, who gave up mathematics because he thought that the only truly important thing in life was to work for the salvation of one's soul. And finally it turned out that the laws of probability are an essential ingredient of the laws of nature. This insight started in the 19th century with Boltzmann and culminated in our century with the development of quantum theory. Einstein could never overcome his intuitive objections against this development. He said: "God does not play dice." Niels Bohr's answer to that was: "We cannot tell God how to shape the universe".

Coming back to the causes of evolution: We can never refute those who say that everything is due to pure chance. At best, we may be able to embarrass them. But it is hard to see here a response to a need. The theory had been developed long before it was needed.

5. WHAT MAKES A MATHEMATICIAN? There exists a widespread resentment against mathematics. It is supposed to deal only with quantity (not true, since most of mathematics deals with structure and relations), or with computing (again not true, but I cannot explain that in a few words) and, on the whole, it is more worthy of a machine than of a human being. As an aid to science and technology, it does not provide values and is therefore dehumanizing. Even the claim of the mathematician to be concerned with truth is frequently answered by saying that mathematical statements are not true but merely correct. Nevertheless, it is undoubtedly true that the results of mathematics are found by human beings. Can anything be said about them?

The answer is: Not enough to enable us to recognize a mathematician if we meet one at a party. Nevertheless, there exist properties without which a mathematician cannot exist. One of them is, of course, a specific talent. But this is far from being enough. It must be supplemented by an interest in the matter, in fact by a fascination with the problems of the field. And the talent must be supported by persistence and by the willingness to spend the large amounts of time and energy needed to master a difficult craft. And the mathematician needs an exceptionally great ability to stand up under frustration. This is due to the fact (pointed out to me by a colleague) that ours is the only field with an all-or-nothing alternative. A painting or a piece of furniture may be more or less perfect. A theorem and a proof are either true or false. If either the proof or the theorem is false, we have absolutely nothing. Finally, we must be satisfied with the production of something intangible. I have found housepainting to be a gratifying supplement to mathematical research. At least one can see and touch what one has done.

It follows that the mathematician needs the support of a civilization that acknowledges as valuable the products of theory, of pure thought. Although we do not set a scale of values, we would not exist without such a scale. I can be brief here, since the arguments given by the philosopher Cantore for the humanistic significance of science apply, with small modifications, to mathematics as well.

Let me conclude by pointing out one advantage that the mathematician (and, with him, the representative of the exact sciences) has. Our thoughts are eminently communicable. Not, perhaps, from person to person. But certainly from nation to nation. Mathematicians understand each other no matter where they come from. Even across many centuries we understand each other. We may not see clearly what a particular expression in Euclid means. But we are confident that, could we talk with him, we would be able to clear up the matter quickly. Nothing is more international than the community of mathematicians.

PROBLEMS AND SOLUTIONS

Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**
with the collaboration of Paul T. Bateman, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Underwood Dudley, Michael A. Filaseta, Ira M. Gessel, Bart Goddard, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Robert Israel, Murray S. Klamkin, Daniel J. Kleitman, Fred Kochman, Frederick W. Luttmann, Frank B. Miles, M. J. Pelling, Richard Pfiefer, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before August 31, 1997; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.*

PROBLEMS

10578. *Proposed by Herbert S. Wilf, University of Pennsylvania, Philadelphia, PA.* Consider the sequence $y_2, y_3, \ldots$ defined by the recurrence relation

$$(n+1)(n-2)y_{n+1} = n(n^2-n-1)y_n - (n-1)^3 y_{n-1}$$

and initial conditions $y_2 = y_3 = 1$. Show that y_n is an integer if and only if n is prime.

10579. *Proposed by Daniel Goffinet, St.-Etienne, France.* Call a function $f : \mathbb{R} \to \mathbb{R}$ *affinely even* if, for some $a \in \mathbb{R}$, $f(a+x) = f(a-x)$ for every $x \in \mathbb{R}$.
(a) Is every function $F : \mathbb{R} \to \mathbb{R}$ the sum of two affinely even functions?
(b) Is every continuous function $F : \mathbb{R} \to \mathbb{R}$ the sum of two continuous affinely even functions?

10580. *Proposed by Stephen C. Locke, Florida Atlantic University, Boca Raton, FL.* Let G be a simple graph with v vertices and e edges and with maximum degree at most 3. Suppose that no component of G is a complete graph on 4 vertices. Prove that G contains a bipartite subgraph H with at least $e - v/3$ edges.

10581. *Proposed by Stephen Herschkorn, RUTCOR, New Brunswick, New Jersey.* Let X be a nonnegative random variable such that the event $\{X > k\}$ has positive probability for every real number k. Consider the collection of nonnegative integer-valued random variables N that are independent of X. Must there exist such an N for which $E(x^N)$ is finite for every real number x but $E(X^N)$ is infinite?

10582. *Proposed by Peter Lindqvist and Kristian Seip, Norwegian University of Science and Technology, Trondheim, Norway.* Let $\mu(n)$ denote the Möbius function and $\zeta(s)$ denote the Riemann zeta function. Prove that

$$\zeta(s)\sum_{m=1}^{N}\sum_{n=1}^{N}\frac{\left(\gcd(m,n)\right)^s}{(mn)^s}\mu(m)\mu(n) = 1 + \sum_{j=2}^{\infty}\frac{1}{j^s}\left(\sum_{\substack{n\mid j\\ n>N}}\mu(n)\right)^2$$

when $s > 1$.

10583. *Proposed by Jacob Lurie, Bethesda, MD.* Let U be a nonempty bounded open set in R^n. For any two points p and q on the boundary of U, suppose there is an affine transformation sending U to itself carrying p to q. Show that there is an affine transformation that carries U to the unit ball.

10584. *Proposed by Charles Conley, Oklahoma State University, Stillwater, OK.* Let $f: \mathbb{R} \to \mathbb{R}$ be a continuous function such that, for all $d \in \mathbb{R}$, the function $f_d: \mathbb{R} \to \mathbb{R}$ defined by $f_d(x) = f(x+d) - f(x)$ is infinitely differentiable. Is f itself infinitely differentiable?

SOLUTIONS

Rational Approximations with Odd Denominators

10242 [1992, 675]. *Proposed by S. Brocco, Brandeis University, Waltham, MA, and F. Mignosi, Institut Blaise Pascal, Paris, France and Università di Palermo, Palermo, Italy.* Let α be a fixed irrational number.

(a) For fixed integer n with $n > 1$, show that it is possible to find a constant $c(n)$ such that there are infinitely many rationals p/q with q relatively prime to n and $|\alpha - p/q| < c(n)/q^2$.

(b) If the continued fraction of α has unbounded partial quotients and $\epsilon > 0$ is given, can one find $c(n) < \epsilon$ satisfying the above condition?

Solution by the editors, based on solutions by the proposers and by the late Raphael M. Robinson. As in G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, fifth ed., Oxford, 1979, we denote the continued fraction for α by $[a_0, a_1, \ldots]$, with the convergents $p_m/q_m = [a_0, a_1, \ldots, a_m]$. The denominators of the convergents satisfy

$$q_0 = 1,\ q_1 = a_1,\ \text{and}\ q_m = a_m q_{m-1} + q_{m-2}. \qquad (*)$$

It follows from $(*)$ and a similar recurrence for the p_m, that $p_m q_{m-1} - q_m p_{m-1} = (-1)^{m-1}$, so that

$$\frac{p_m}{q_m} - \frac{p_{m-1}}{q_{m-1}} = \frac{(-1)^{m-1}}{q_m q_{m-1}}.$$

This leads to estimates on $|\alpha - p_m/q_m|$. In particular, it is known that every convergent satisfies $|\alpha - p_m/q_m| < 1/q_m^2$; while if $|\alpha - p/q| < 1/(2q^2)$, then $p/q = p_m/q_m$ for some m.

Consider first the case where n is prime. Then $(*)$ shows that $\gcd(q_m, q_{m-1}) = 1$ for $m = 1, 2, \ldots$, so at least one of the pair $\{q_m, q_{m-1}\}$ must be relatively prime to n. Infinitely many convergents thus satisfy the required property, so we may take $c(n) = 1$ whenever n is prime.

In general, one cannot expect that infinitely many q_m are relatively prime to a composite n. For example, take $n = 6, a_0 = 0, a_1 = 2, a_2 = 1$, with $2|a_{2k-1}$ and $3|a_{2k}$ for $k > 1$. Then $(*)$ shows that $2|q_{2k-1}$ and $3|q_{2k}$ for all $k \geq 1$. In particular, this gives a negative answer to **(b)**, since the a_m may grow rapidly, while $c(6) > 1/2$ since $\gcd(q, 6) = 1$ requires that p/q is not a convergent.

We now show that we may take $c(n) = (n+1)(n+2)$ in **(a)**. Let $p = bp_m + p_{m-1}$ and $q = bq_m + q_{m-1}$ for a positive integer b. Then $bq_m \leq q \leq (b+1)q_m$, and

$$\left|\alpha - \frac{p}{q}\right| \leq \left|\alpha - \frac{p_m}{q_m}\right| + \left|\frac{p_m}{q_m} - \frac{p}{q}\right| \leq \frac{1}{q_m^2} + \frac{1}{q_m q} \leq \frac{(b+1)^2 + (b+1)}{q^2}.$$

Let b equal the product of those primes that divide n but not q_{m-1}. Then $b \leq n$, and since each prime factor of n divides precisely one of either bq_m or q_{m-1}, it cannot divide q. Thus q is relatively prime to n.

A negative answer to (**b**) for any given n can be found by choosing α with $a_m = n$ if m is odd, while the $a_m \to \infty$ if m is even. Then $(*)$ shows that $q_m \equiv 0 \pmod{n}$ if m is odd and $q_m \equiv 1 \pmod{n}$ if m is even. However, for even m, $|\alpha - p_m/q_m| > 1/\big((n+2)q_m^2\big)$.

Editorial comment. Much more precise information relating the approximation of α by the convergents p_m/q_m to the a_m can be found in T. W. Cusick and M. Flahive, *The Markoff and Lagrange Spectra*, American Mathematical Society, Providence, 1989. From this point of view, one should fix n and consider how the quantity here called $c(n)$ depends on α. Some special cases are known. For example, the case $n = 2$ was studied in Raphael M. Robinson, The approximation of irrational numbers by fractions with odd or even term, *Duke Math. J.* 7 (1940), 354–359. There it is shown that $|\alpha - p/q| < 1/(2q^2)$ always has infinitely many solutions with q odd, and the numbers α for which this is best possible are characterized.

Two Klein Bottles and a Torus

10359 [1994, 76]. *Proposed by Raphael M. Robinson, University of California, Berkeley, CA.* Two pairs of sides of the unit square $0 \le x \le 1, 0 \le y \le 1$ are identified in such a way that the surface obtained has a locally Euclidean metric. How many such surfaces are there that are inequivalent as metric spaces?

Solution by the proposer. There are three such manifolds. The identifications must be made in such a way that all four vertices are identified, in order to have a Euclidean metric near a vertex. There are three essentially different ways of doing this, which lead to a torus T and Klein bottles K and L. The following identifications are to be made for $0 \le t \le 1$: T identifies $(t, 0)$ with $(t, 1)$ and $(0, t)$ with $(1, t)$; K identifies $(t, 0)$ with $(t, 1)$ and $(0, t)$ with $(1, 1-t)$; and L identifies $(t, 0)$ with $(1, t)$ and $(0, t)$ with $(t, 1)$.

The torus T is topologically different from K and L, since T is orientable and K and L are not. Although K and L are topologically equivalent, they are inequivalent as metric spaces. One difference is that shortest segments in K whose endpoints coincide have length 1, whereas in L there are such segments of length $\sqrt{2}/2$, for example the one joining $(1/2, 0)$ to $(1, 1/2)$.

Remark. Let $K(a, b)$ be the manifold obtained by starting with the rectangle $0 \le x \le a$, $0 \le y \le b$ and identifying $(t, 0)$ with (t, b) for $0 \le t \le a$, and $(0, t)$ with $(a, b-t)$ for $0 \le t \le b$. Then $K = K(1, 1)$, and it can be shown that L is equivalent to $K(\sqrt{2}/2, \sqrt{2})$ as a metric space. All of the manifolds $K(a, b)$ are topologically equivalent, but no two of them are equivalent as metric spaces.

A False Leaky Tent

10367 [1994, 176]. *Proposed by Donald R. Chalice, Western Washington University, Bellingham, WA.* Let $\mathbf{C}$ be the Cantor set in [0, 1], and let $\mathbf{E}$ be the set of endpoints of the removed intervals (together with 0 and 1). Let $\mathbf{F} = \mathbf{C} - \mathbf{E}$ and let p be the point $(1/2, 1/2)$. For any $c \in \mathbf{C}$, let L_c be the line segment from p to c, let Q_c be the points on L_c with rational ordinates and I_c the points of L_c with irrational ordinates. The set

$$\mathbf{T} = \left(\bigcup_{e \in \mathbf{E}} Q_e\right) \cup \left(\bigcup_{f \in \mathbf{F}} I_f\right)$$

has the property that $\mathbf{T}$ is connected, but $\mathbf{T} - p$ is totally disconnected. Consider instead

$$\mathbf{T}_0 = \left(\bigcup_{e \in \mathbf{E}} I_e\right) \cup \left(\bigcup_{f \in \mathbf{F}} Q_f\right)$$

obtained by interchanging the roles of points with rational and irrational ordinates. Is $\mathbf{T}_0$ connected?

Solution by Kenneth Schilling, University of Michigan, Flint, MI. No, it is not. It suffices to construct a continuous function ϕ from $\mathbf{C}$ to the open interval $(0, 1/2)$ such that $\phi(c)$ is rational for $c \in \mathbf{E}$ and irrational for $c \in \mathbf{F}$; then the sets

$$\mathbf{T}_0 \cap \{ (x, y) \in L_c : y < \phi(c), c \in \mathbf{C} \} \quad \text{and} \quad \mathbf{T}_0 \cap \{ (x, y) \in L_c : y > \phi(c), c \in \mathbf{C} \}$$

are disjoint, open relative to $\mathbf{T}_0$, and cover $\mathbf{T}_0$, and so $\mathbf{T}_0$ is not connected.

To construct such a ϕ, say $c = .c_1c_2c_3c_4 \ldots \in \mathbf{C}$, expressed in ternary digits using only the digits 0 and 2. Put $\phi(c) = y = .0y_2y_3y_4 \ldots$, where $y_n = 1$ if $c_n = c_{n-1}$, and y_n is, say, the kth digit of π if c_n is the kth digit of c that differs from its predecessor. The idea here is that if $c \in \mathbf{E}$ and so the ternary digit of c is eventually constant, then so is the representation of $\phi(c)$, so $\phi(c)$ is rational. If $c \in \mathbf{F}$ and so the representation of c is not eventually constant, then the representation of $\phi(c)$ is not eventually repeating, so $\phi(c)$ is irrational. It is easy to check that ϕ is continuous.

Editorial comment. All solvers used an argument similar to that of the selected solution. John Cobb noted that the *middle thirds* set implied by the phrase "*the* Cantor set" is only one of a huge family of homeomorphic sets. He constructs a set $\widetilde{\mathbf{C}}$ in this family such that: (1) every endpoint is rational, (2) every non-endpoint is irrational, and (3) $1/2 \notin \widetilde{\mathbf{C}}$. If $\widetilde{\mathbf{T}}_0$ is constructed from this set using $(1/2, 1/2)$, as in the problem statement, then every vertical line $x = a$ where a is rational and not in $\widetilde{\mathbf{C}}$ separates $\widetilde{\mathbf{T}}_0$.

K. P. Hart gave two solutions. One resembled the selected solution, while the other showed that $\mathbf{T}_0$ is *zero-dimensional*; that is, it has a base for its topology that consists of closed-and-open sets. The sets $I_e \cup \{(1/2, 1/2)\}$ for $e \in E$ and $V_q = \{ (x, y) \in \mathbf{T}_0 : y = q \}$ form a countable family of closed zero-dimensional subspaces of $\mathbf{T}_0$ whose union is $\mathbf{T}_0$. The conclusion then follows from the Countable Closed Sum Theorem of Dimension Theory (see R. Engelking, *Dimension Theory*, North-Holland, 1978, Theorem 1.3.1).

Solved also by J. Cobb, R. Griffus, K. P. Hart (The Netherlands), O. P. Lossers (The Netherlands), M. D. Meyerson, A. W. Schurle, and the proposer.

Reverse Dynamics of a Polynomial Map on the Line

10369 [1994, 273]. *Proposed by Bjorn Poonen (student), University of California, Berkeley, CA.* Let $f(x)$ be a polynomial having rational coefficients and degree at least 2. Suppose that $a_1, a_2, a_3, \ldots$ is a sequence of rational numbers such that $f(a_{n+1}) = a_n$ for all $n \geq 1$. Prove that there exists $k \geq 1$ such that $a_{n+k} = a_n$ for all $n \geq 1$.

Solution by Emre Alkan (student), Bosphorus University, Istanbul, Turkey. We first show that $\{a_n\}$ is bounded. Clearly $\lim_{x\to\infty} |f(x)|/|x| = \infty$, so there exists a constant M such that $M \geq |a_1|$ and $|f(x)| \geq |x|$ whenever $|x| \geq M$. If $|a_n| > M$, then $|a_{n-1}| = |f(a_n)| \geq |a_n| > M$. Repeating this yields $|a_1| > M$, a contradiction. Hence $|a_n| \leq M$ for all n.

We next construct an integer N such that each Na_n is an integer. Express f by $f(x) = (b_dx^d + b_{d-1}x^{d-1} + \cdots + b_0)/c$, where $c, b_o, b_1, \ldots, b_d$ are integers. Let r, s be integers such that $a_1 = r/s$. Let $N = sb_d$. Clearly $Na_1 = rb_d$ is an integer. If Na_n is an integer, then $(cN^d/b_d)(f(x/N) - a_n)$ is a monic polynomial with integer coefficients that vanishes at Na_{n+1}. By the Rational Zeros Theorem (G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, fifth ed., Oxford, 1979, Theorem 45, section 4.3, p. 41), Na_{n+1} must also be an integer. By induction, Na_n is an integer for all n.

The set $\{a_n\}$ is bounded, and its elements are multiples of $1/N$. The set is therefore finite; let m be its size. Consider segments consisting of $m + 1$ consecutive elements of the sequence. Each such segment has one of m^{m+1} patterns, and each pattern has a repeated element. Since there are infinitely many segments and finitely many patterns, some pattern appears infinitely often.

This yields a repetition—with some fixed displacement k—that occurs infinitely many times in the sequence. In particular, for each $n \geq 1$ there exists $j > n$ such that $a_j = a_{j+k}$. Repeatedly applying f to both sides of this equation yields $a_n = a_{n+k}$, as desired.

Solved also by R. Barbara (Lebanon), P. Budney, R. Holzsager, R. B. Israel (Canada), J. H. Lindsey II, O. P. Lossers (The Netherlands), M. Reid, R. M. Robinson, G. L. Stanek, J. Sturm, A. N. 't Woord (The Netherlands), and the proposer.

Hamiltonian Paths Through Isomorphism Classes of Graphs

10370 [1994, 273]. *Proposed by Bernardo Recamán Santos, Waterford Kamhlaba United World College of South Africa, Mbabane, Swaziland.* Let G_n be the undirected graph whose vertices are the unlabeled graphs on n vertices (e.g. G_4 has 11 vertices), two of which are adjacent in G_n if and only if one can be obtained from the other by deleting an edge.
(**a**) Show that neither G_4 nor G_5 contains a Hamiltonian path.
(**b**)* Does G_n contain a Hamiltonian path for $n > 5$?

Solution by Frank Schmidt, Arlington, VA.

(**a**) The graph G_4 is shown in Figure 10370.

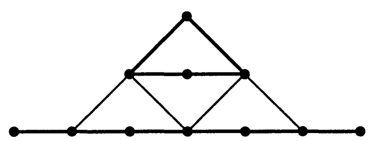


Figure 10370

Suppose G_4 has a Hamiltonian path. Since G_4 has 11 vertices, the path contains 10 edges. The endpoints of the path must be the vertices of degree 1, and hence the path must contain all edges incident to vertices of degree at most 2. The ten such edges in G_4 (shown in heavier lines in Figure 10370) do not form a path, so G_4 has no Hamiltonian path. It follows from the partial solution of part (**b**) that G_5 does not have a Hamiltonian path.

(**b**) (Partial solution) If $n \geq 5$ and n is congruent to 0 or 1 modulo 4, then G_n has no Hamiltonian path. A Hamiltonian path in G_n must alternate between graphs of even size and graphs of odd size. This requires $|e(n) - o(n)| \leq 1$, where $e(n)$ and $o(n)$ denote the number of isomorphism classes of n-vertex graphs with an even or an odd number of edges, respectively. In the solution of MONTHLY Problem 10285 [1993, 185; 1996, 268], it was noted that $e(n) - o(n) = s(n)$, where $s(n)$ is the number of isomorphism classes of self-complementary graphs with n vertices. This contradicts the known result that $s(n) \geq 2$ when $n \geq 5$ and n is congruent to 0 or 1 modulo 4 (see M. Kropar and R. Reed, On the construction of self-complementary graphs on 12 nodes, *J. Graph Theory*, 3 (1979), 111–125).

Editorial comment. When n is congruent to 2 or 3 modulo 4, there are no self-complementary n-vertex graphs, and the partite sets of G_n are equinumerous. This may permit a Hamiltonian path. In particular, G_2 and G_3 are paths. It would be of interest to study this question for G_6, which has 156 vertices and 572 edges.

A self-contained proof of the above results was given by A. N. 't Woord (The Netherlands); S. C. Locke solved part (a) and the cases n=8 and n=9 of part (b); while G. Laman (The Netherlands) and the proposer solved only part (a).

Maximizing a Ratio of Areas

10371 [1994, 274]. *Proposed by Emil Yankov Stoyanov, Antiem I Mathematical School, Vidin, Bulgaria.* Let B' and C' be points on the sides AB and AC, respectively, of a given triangle ABC, and let P be a point on the segment $B'C'$. Determine the maximum value of

$$\frac{\min\{[BPB'], [CPC']\}}{[ABC]}$$

where [**F**] denotes the area of **F**.

Solution by Victor Pambuccian, Arizona State University West, Phoenix, AZ. This is a problem of ordered affine geometry, since it asks for the maximum value of

$$\min\left\{\frac{[BPB']}{[ABC]}, \frac{[CPC']}{[ABC]}\right\},$$

and ratios of areas of triangles are affine notions (see Léonce Lesieur, Sur la mesure des triangles en géométrie affine, *Math. Z.* 93 (1966), 334–344).

Whenever X' is a point on the side YZ of a triangle XYZ, we have $[XX'Y]/[XYZ] = |YX'|/|YZ|$. Using this, we get

$$\frac{[BB'P]}{[BB'C']} = \frac{|B'P|}{|B'C'|}, \quad \frac{[BB'C']}{[ABC']} = \frac{|BB'|}{|BA|}, \quad \frac{[ABC']}{[ABC]} = \frac{|AC'|}{|AC|}.$$

Hence

$$\frac{[BB'P]}{[ABC]} = \frac{|B'P|}{|B'C'|} \cdot \frac{|BB'|}{|BA|} \cdot \frac{|AC'|}{|AC|}, \tag{1}$$

and analogously

$$\frac{[CC'P]}{[ABC]} = \frac{|C'P|}{|C'B'|} \cdot \frac{|CC'|}{|CA|} \cdot \frac{|AB'|}{|AB|}. \tag{2}$$

Multiplying the right sides of (1) and (2) and grouping the factors conveniently, we get

$$\left(\frac{|B'P|}{|B'C'|} \cdot \frac{|C'P|}{|C'B'|}\right) \cdot \left(\frac{|BB'|}{|BA|} \cdot \frac{|AB'|}{|AB|}\right) \cdot \left(\frac{|AC'|}{|AC|} \cdot \frac{|CC'|}{|CA|}\right).$$

The two factors inside each pair of parentheses add up to 1, hence their product is at most $1/4$, with equality if and only if the respective terms are equal. Therefore the whole product is at most $(1/4)^3 = (1/8)^2$, with equality if and only if B', C', and P are midpoints of AB, AC, and $B'C'$, respectively.

Since the product of $[BPB']/[ABC]$ and $[CPC']/[ABC]$ is always at most $(1/8)^2$, the smaller of the two is at most $1/8$, with equality if and only if B', C', and P are midpoints of AB, AC, and $B'C'$, respectively. Hence the required maximum is $1/8$.

Solved also by J. Anglesio (France), R. Barbara (Lebanon), K. L. Bernstein, R. J. Chapman (U. K.), J.-P. Grivaux (France), N. Komanda, J. H. Lindsey II, O. P. Lossers (The Netherlands), C. G. Petalas (Greece), R. Reynolds, D. Tang, A. N. 't Woord (The Netherlands), NSA Problems Group, and the proposer.

A Recurrence with a Harmonic Solution

10375 [1994, 362]. *Proposed by John Brillhart and J. S. Lomont, University of Arizona, Tucson, AZ.* Find the complete solution of the recurrence

$$U_{n+2} = 2(2n+3)^2 U_{n+1} - 4(n+1)^2(2n+1)(2n+3)U_n \qquad (n \geq 0).$$

Solution by Kee-Wai Lau, Hong Kong. We show that for $n \geq 1$,

$$U_n = (2n)!\big(U_0 + (U_1/2 - U_0)(1 + 1/2 + 1/3 + \cdots + 1/n)\big).$$

The substitution $V_n = U_n/(2n)!$ yields

$$(n+2)V_{n+2} = (2n+3)V_{n+1} - (n+1)V_n$$

for $n \geq 0$. This is equivalent to $(n+2)(V_{n+2} - V_{n+1}) = (n+1)(V_{n+1} - V_n)$, which implies that $(n+2)(V_{n+2} - V_{n+1}) = V_1 - V_0$. Rewriting this as $V_{n+2} = V_{n+1} + (V_1 - V_0)/(n+2)$ yields the solution

$$V_n = V_0 + (V_1 - V_0)(1 + 1/2 + \cdots + 1/n),$$

which translates into the desired formula.

Editorial comment. Several solvers began by observing that $U_n = (2n)!$ is a solution. The second solution can then be found by reduction of order. In particular, Marko Petkovšek claimed to find that solution "by inspection" although he indicated in a footnote to his solution that the inspection was done by a computer. See M. Petkovšek, Hypergeometric solutions of linear recurrences with polynomial coefficients, *J. Symb. Comp.* 14 (1992), 243–264 for more information about this approach.

Solved also by J. Alvarez (Spain), J. Anglesio (France), R. Bagby, H. J Barten, K. L. Bernstein, J. C. Binz (Switzerland, two solutions), G. A. Bookhout, M. Burger (Austria), R. J. Chapman (U. K.), C. K. Cook, P. Deiermann, D. Doster, S. B. Ekhad, C. Georghiou (Greece), J.-P. Grivaux (France), C. R. Hampton, M. S. Klamkin (Canada), A. M. Krall, J. Laforgue, K.-W. Lau (Hong Kong), G. Letac (France), J. H. van Lint (The Netherlands), S. C. Locke, O. P. Lossers (The Netherlands), G. Loudner, C. Mallinger (Austria), L. E. Mattics, J. Ottenstein (Israel), A. Pedersen (Denmark), M. Petkovšek (Slovenia), R. C. Read (Canada), R. Richberg (Germany), N. C. Singer, M. Vowe (Switzerland), H. Widmer (Switzerland), Anchorage Math Solutions Group, NSA Problems Group, and the proposers.

A Diophantine Polynomial Equation

10376 [1994, 362]. *Proposed by Nobuhisa Abe, Oita, Japan.* Determine all integer solutions of

$$x(x+1)(x+2)(x+3)(x+4)(x+5) = y^2 - 1.$$

Solution by Paul T. Bateman, University of Illinois, Urbana, IL. Aside from the trivial solutions given by $x = -5, -4, -3, -2, -1, 0$ and $y = \pm 1$, we claim that the only integer solutions of the given equation are $(x, y) = (-7, \pm 71)$ or $(2, \pm 71)$.

Let $f(x) = x(x+1)(x+2)(x+3)(x+4)(x+5) + 1$. We seek those integer values of x such that $f(x)$ is a square or, equivalently, $64f(x)$ is a square. Since $f(-x) = f(x-5)$, we may restrict attention to positive integer values of x. For large positive x we have the following expansion in decreasing powers of x:

$$\begin{aligned}\left(64f(x)\right)^{1/2} &= 8x^3\left((1+x^{-1})(1+2x^{-1})(1+3x^{-1})(1+4x^{-1})(1+5x^{-1}) + x^{-6}\right)^{1/2} \\ &= 8x^3\left(1 + 15x^{-1} + 85x^{-2} + 225x^{-3} + O(x^{-4})\right)^{1/2} \\ &= 8x^3 + 60x^2 + 115x + 75/2 + O(x^{-1}).\end{aligned}$$

Hence for large positive x we have

$$(8x^3 + 60x^2 + 115x + 37)^2 < 64f(x) < (8x^3 + 60x^2 + 115x + 38)^2. \qquad (*)$$

A brief calculation gives

$$(8x^3 + 60x^2 + 115x + 38)^2 - 64f(x) = 8x^3 + 249x^2 + 1060x + 1380$$

and

$$64f(x) - (8x^3 + 60x^2 + 115x + 37)^2 = 8x^3 - 129x^2 - 830x - 1305.$$

Hence the right-hand inequality in $(*)$ holds for all positive x and the left-hand inequality in $(*)$ holds for $x \geq 22$. Thus, if x is any integer greater than 21, $(*)$ shows that $64f(x)$ lies strictly between two consecutive squares and so is not a square. On the other hand, it is easy to verify (using a pocket calculator, say) that $f(x)$ is not a square for $x = 1$ and for $x = 3, 4, 5, \ldots, 21$. Thus our claim is established.

This method goes back to C. Runge, Über ganzzahlige Lösungen von Gleichungen zwischen zwei Veränderlichen, *J. Reine Angew. Math.* 100 (1887), 425–435.

Editorial comment. L. E. Mattics made a systematic study of the equation with $y^2 - k$ in place of $y^2 - 1$ on the right side, and found that the positive solutions with $|k| \leq 31$ are: (3, 142) for $k = 4$, (1, 27) for $k = 9$, and (21, 12875) for $k = 25$. He proved also that for positive solutions, $1 \leq x \leq \max\{27, (40\,|k| + 1)^{1/3}\}$.

Solved also by Y. Alemu (Ethiopia), J. Anglesio (France), D. Caccia, M. J. Cohen, H. G. Killingbergtrø (Norway), N. Komanda, J. H. Lindsey II, S. C. Locke & A. D. Meyerowitz, O. P. Lossers (The Netherlands), L. E. Mattics, M. Reid, J. P. Robertson, R. M. Robinson, F. Schmidt, N. C. Singer, A. N. 't Woord (The Netherlands), Anchorage Math Solutions Group, and the proposer.

Hermitian Matrices

10377 [1994, 362]. *Proposed by Kathryn R. Laberteaux (student), University of Michigan, Ann Arbor, MI.* On the final exam in a linear algebra class, I was asked to express the statement "A is Hermitian" in the form of a matrix identity. I should have written "$A = A^*$", but out of haste and exhaustion I wrote "$AA^* = A^2$" instead. Was my answer correct?

Solution I by Marvin Marcus, Santa Barbara, CA. Yes. By the Schur Triangularization Theorem, we may write $A = U^*TU$, where U is unitary and T is upper triangular with diagonal $\lambda_1, \ldots, \lambda_n$ being the eigenvalues of A. Then $AA^* = A^2$ becomes $TT^* = T^2$. Now $\mathrm{tr}(TT^*) = \mathrm{tr}(T^2)$ yields $\sum_{k=1}^n |\lambda_k|^2 + \sum_{1\le i<j\le n} |t_{ij}|^2 = \sum_{k=1}^n \lambda_k^2$. The triangle inequality implies that $t_{ij} = 0$ for all i, j. We then have $|\lambda_k|^2 = \lambda_k^2$ for all k, and hence each λ_k is real. In other words, A is unitarily similar to a diagonal matrix with real entries and thus is Hermitian.

Solution II by G. P. Shannon, University of Ulster, Coleraine, N. Ireland. To show that $A - A^* = 0$, it suffices to show that $\mathrm{tr}\left((A - A^*)(A - A^*)^*\right) = 0$. We first compute $(A-A^*)(A-A^*)^* = (A-A^*)(A^*-A) = (AA^*-A^2)+(A^*A-A^{*2})$. Using the linearity of the trace function, the property $\mathrm{tr}(AA^*) = \mathrm{tr}(A^*A)$, and the condition $AA^* = A^2 = (A^*)^2$, we obtain $\mathrm{tr}\left((A - A^*)(A - A^*)^*\right) = \mathrm{tr}(A^*A) - \mathrm{tr}(A^{*2}) + \mathrm{tr}(AA^*) - \mathrm{tr}(A^2) = 0$, and so A is Hermitian.

Solution III by Eugene A. Herman, Grinnell College, Grinnell, IA. Yes, the equations $A^* = A$ and $AA^* = AA$ are equivalent. Multiplying the first by A yields the second. To derive the first from the second, we use that the null space of A and the range of A^* are orthogonal complements with respect to complex inner product. Thus $A^* = A$ follows by proving that $A^*v = Av$ for all $v \in N(A)$ and all $v \in R(A^*)$. If $v \in N(A)$, then $Av = 0$, and we compute $||A^*v||^2 = \langle A^*v, A^*v\rangle = \langle v, AA^*v\rangle = \langle v, AAv\rangle = \langle v, 0\rangle = 0$. This yields $A^*v = 0$, and hence $A^*v = Av$. When $v = A^*w \in R(A^*)$, we use the complex conjugate $AA^* = A^*A^*$ of $AA^* = AA$ to obtain $A^*v = A^*A^*w = AA^*w = Av$, which completes the proof.

Solution IV by Robert B. Israel, University of British Columbia, Vancouver B. C., Canada. Assume that $AA^* = A^2$. We first prove by induction that $A^n = (A^*)^n$ for $n \ge 2$. Conjugating the given equation yields this for $n = 2$. If $A^n = (A^*)^n$ for $n \ge 2$, then $A^{n+1} = A(A^*)^n = (AA^*)(A^*)^{n-1} = (A^*)^2(A^*)^{n-1} = (A^*)^{n+1}$.

Let p be the monic minimal polynomial of A. Every eigenvalue of $A^2 = AA^*$ is a nonnegative real number, so every eigenvalue of A is real. Therefore p has real coefficients and $p(A^*) = (p(A))^* = 0$. Thus p is also the minimal polynomial of A^*.

If $(A^*)^2v = 0$, that is $AA^*v = 0$, then $A^*v = 0$. Thus t occurs at most once as a factor of $p(t)$, and the linear coefficient a_1 or constant coefficient a_0 in $p(t)$ is nonzero. If $a_0 = 0$, then $0 = p(A) - p(A^*) = a_1(A - A^*)$ with $a_1 \ne 0$, so $A = A^*$. If $a_0 \ne 0$, then $0 = Ap(A) - A^*p(A^*) = a_0(A - A^*)$ and again $A = A^*$.

Solution V by Gérard Letac, Université Paul Sabatier, Toulouse, France. Consider the Hermitian matrix $B = i(A-A^*)$. The hypothesis $AA^* = A^2$ yields $AB = 0$. If B has a nonzero eigenvalue λ with eigenvector x, then $0 = ABx = \lambda Ax$. Thus $0 = Ax = A^*x - i\lambda x$, and $0 = x^*Ax = x^*A^*x - i\lambda x^*x$. Since $x^*Ax = \overline{x^*A^*x}$, we obtain $\lambda = 0$, a contradiction. Since the eigenvalues of the Hermitian matrix B are zero, $B = 0$.

Solved also by A. Bandopadhyay (India), S. K. Berberian, R. Bielawski (Canada), P. Budney, D. Callan, R. J. Chapman (U. K.), M.-D. Choi (Canada) & C.-K. Li & W. So, D. Choudhury, L. M. DeAlba, G. Ehrlich, D. Fasino (Italy), P. M. Gibson, S. Goldberg, J.-P. Grivaux (France), J. L. Hartman, A. L. Holshouser & B. G. Klein, R. A. Horn, N. Komanda, D. W. Koster, K. Kubo (Japan), C. Lanski, J. H. Lindsey II, O. P. Lossers (The Netherlands), W. Margulies, T. L. Markham & A. R. Schep, J. K Merikoski & A. Virtanen (Finland), J. M. Monier (France), J. Rätz (Switzerland), R. Richberg (Germany), F. Richman, F. Schmidt, A. Singh (India), G. Trenkler (Germany), E. I. Verriest (France), A. N. 't Woord (The Netherlands), D. J. Wright, P. Y. Wu (China), F. Zhang, M. Zhang (China), the New Mexico Tech Problem Solving Group, and the proposer.

From Integers to Integers, But Not Very Many

10382 [1994, 473]. *Proposed by Richard K. Guy, University of Calgary, Calgary, Alberta, Canada.* Which integers are represented by $(x+y+z)^2/(xyz)$ where x, y, and z are positive integers?

Solution by Rolf Richberg, RWTH Aachen, Aachen, Germany. The numbers so represented are $\{1, 2, 3, 4, 5, 6, 8, 9\}$.

Let $F(x, y, z) = (x+y+z)^2/(xyz)$. Fix n and suppose $n = F(x, y, z)$, with $x \le y \le z$ and z minimal for that choice of n. From

$$nxyz = (x+y+z)^2 = (x+y)^2 + 2(x+y)z + z^2,$$

we infer that $z|(x+y)^2$. If $z > x+y$, then $(x+y)^2/z < z$ and

$$F\left(x, y, \frac{(x+y)^2}{z}\right) = \frac{\frac{(x+y)^2}{z^2}(x+y+z)^2}{xy\frac{(x+y)^2}{z}} = \frac{(x+y+z)^2}{xyz} = n.$$

Thus the minimality of z implies that $x+y \ge z$. Now

$$n = \frac{x}{yz} + \frac{y}{xz} + \frac{z}{xy} + \frac{2}{x} + \frac{2}{y} + \frac{2}{z} \le \frac{1}{z} + \frac{1}{x} + \left(\frac{1}{y} + \frac{1}{x}\right) + \frac{2}{x} + \frac{2}{y} + \frac{2}{z} \le \frac{7}{x} + \frac{3}{z}.$$

This implies that $z = 1$ (and $n = 9$) or that $z \ge 2$ (and $n \le 8$). Thus $n \le 9$.

We next prove that $n \ne 7$. The inequality $7 \le 7/x + 3/z$ prohibits $x \ge 2$. With $x = 1$, $x + y \ge z$ yields $y \le z \le y+1$. When $z = y$ we have $(1+2y)^2 = 7y^2$, and when $z = y+1$ we have $(2+2y)^2 = 7y(y+1)$, neither of which has an integer solution.

Finally, $F(9,9,9) = 1$, $F(4,4,8) = 2$, $F(3,3,3) = 3$, $F(2,2,4) = 4$, $F(1,4,5) = 5$, $F(1,2,3) = 6$, $F(1,1,2) = 8$, and $F(1,1,1) = 9$.

Solved also by J. Anglesio (France), M. Barr (Canada), B. Battsengel (Mongolia), R. J. Chapman (U. K.), J. Christopher, M. Farris, D. Khaykis, P. A. Kumar & V. Kannan (India), G. N. Lewis, J. H. Lindsey II, L. E. Mattics, J. McHugh, R. E. Prather, J. P. Robertson & J. S. Robertson, R. M. Robinson, N. C. Singer, D. West, NSA Problems Group, and the proposer.

Spaces with Closed Fixed-Point Sets

10385 [1994, 474]. *Proposed by Nándor Sieben, Arizona State University, Tempe, AZ.* Let $\mathbf{X}$ be a topological space. It is easy to see that if $\mathbf{X}$ is a Hausdorff space, then *fixed-point sets are closed.* That is, for any continuous function $f: \mathbf{X} \to \mathbf{X}$, the set $F = \{x \in \mathbf{X} : f(x) = x\}$ is closed. Is the converse true? More precisely, if $\mathbf{X}$ has the property that all fixed-point sets are closed, must $\mathbf{X}$ be a Hausdorff space?

Solution by Paul R. Meyer, Lehman College, CUNY, Bronx, NY. Let T_{fix} denote the stated property. One can readily verify that $T_2 \implies T_{\text{fix}} \implies T_1$ and that the cofinite topology on an infinite set shows that the second implication is not reversible.

To give a negative answer to the question in this problem by showing that the first implication is not reversible,we use an example from Helen F. Cullen, Unique sequential limits, *Boll. Un. Mat. Ital.* (3) 20 (1965), 123–124. Let $X = \mathbb{R} \cup \{p\}$ where $p \notin \mathbb{R}$. A set not containing p is open if and only if it is open in the usual topology of $\mathbb{R}$, and a set

containing p is open if and only if its complement is the union of finitely many convergent sequences in $\mathbb{R}$ together with their limits. Clearly X is not T_2 (the set $[0, 1]$ is compact but not closed). To show that X is T_{fix}, let f be a continuous self-map of X and show that its set of fixed points F is closed. There are two cases.

Case 1. $f(p) = p$. To show that F is closed, assume that x is an accumulation point of F with $x \neq p$. Then $x \in \mathbb{R}$ and there is a sequence (x_n) of elements of F converging to x. Since $f(x_n) = x_n$, it follows from the continuity of f that $f(x) = x$. Thus, $x \in F$ and F is closed.

Case 2. $f(p) \neq p$. We complete the proof by showing that f is a constant function. Say $f(p) = 0$ for definiteness. Then $X - f^{-1}(-1/n, 1/n)$ is a countable set, call it A_n. Then $\bigcup A_n = X - f^{-1}(0)$ is also countable, so that $f^{-1}(0)$ is dense in $\mathbb{R}$ by the Baire category theorem. But $f^{-1}(0)$ is also closed in X. It follows that f is constant. Thus $F = \{0\}$, which is closed.

Editorial comment. Joachim Schröder supplied a reference to Vera Trnkova, Categorical aspects are useful for topology, *General Topology and its Relations to Modern Analysis and Algebra (Lecture Notes in Mathematics 609)*, Springer, 1977, where it is shown that every T_1 space V can be embedded as a closed subspace of a T_1 space X such that all continuous self-maps of X are either constant or the identity.

Solved also by B. Burdick, J. Cobb, R. Griffus & K. Smith, A. Kumar & P. Chatterjee (India), R. P. Millspaugh, J. Schröder (South Africa), J. R. Smith, and the proposer.

Accumulated Preferences

10397 [1994, 681]. *Proposed by Sam Northshield, SUNY, Plattsburgh, NY, and José Luis Palacios, Universidad Simón Bolívar, Caracas, Venezuela.* Let X be an $\mathbb{N}$-valued random variable. Show that if $Pr(X = k | X = k \text{ or } X = k+1)$ is non-decreasing, then $Pr(X = k | X \leq k)$ is non-increasing.

Solution by Richard Holzsager, The American University, Washington, DC. Write $p_k = Pr(X = k)$. The hypothesis implies that if $p_k > 0$, then $p_{k+1} > 0$. It won't hurt to assume that the first k for which $p_k > 0$ is $k = 0$ (it's just a matter of renumbering, and saves the trouble of interpreting $Pr(X = 0 | X \leq 0)$ when both events are impossible).

The hypothesis $p_{k-1}/(p_{k-1} + p_k) \leq p_k/(p_k + p_{k+1})$ is equivalent to $p_{k-1}/p_k \leq p_k/p_{k+1}$. By transitivity, $p_i/p_{i+1} \leq p_n/p_{n+1}$, or $p_i/p_n \leq p_{i+1}/p_{n+1}$ for any $i \leq n$, so

$$\frac{p_0 + \cdots + p_n}{p_n} \leq \frac{p_1 + \cdots + p_{n+1}}{p_{n+1}} < \frac{p_0 + \cdots + p_{n+1}}{p_{n+1}}.$$

Inverting, we get the desired result, strengthened to say that the probabilities in question are actually strictly decreasing.

Solved also by A. Adler, S. V. Amari & R. B. Misra (India), C. Anderson, M. H. Andreoli, N. Bouzar, D. D. Brics (Denmark) & D. Ranjan, A. E. Caicedo Núñez (Colombia), D. Callan, R. J. Chapman (U. K.), M. P. Cohen, D. A. Darling, R. Ehrenborg (Canada), N. Grivaux (France), V. Hernández (Spain), E. Hertz, R. D. Hurwitz, R. S. Katti & S. Vidyashankara, F. Kemp, G. Keselman, J. Kupka (Australia), J. H. Lindsey II, O. P. Lossers (The Netherlands), J. Marengo, D. K. Nester, A. Pedersen (Denmark), G. S. Rogers, D. M. Rosenblum, R. P. Sealy (Canada), N. C. Singer, R. Stong, D. B. Tyler, E. I. Verriest (France), E. A. Weinstein, NSA Problems Group, Western Maryland College Problems Group, and the proposers.

A Way to Form Ideals of Power Series

10399 [1994, 682]. *Proposed by Mowaffaq Hajja, Yarmouk University, Irbid, Jordan.* Find all infinite sequences $\mathbf{c} = \langle c_0, c_1, c_2, \ldots \rangle$ of integers for which the set

$$I_{\mathbf{c}} = \left\{ \sum a_i x^i \in \mathbb{Z}[x] : \sum a_i c_i = 0 \right\}$$

is an ideal of $\mathbb{Z}[x]$.

Composite solution by John H. Lindsey II, Ft. Myers, FL and Nasha Komanda, Central Michigan University, Mt. Pleasant, MI. We prove that $I_{\mathbf{c}}$ is an ideal if and only if there are integers r and s such that $c_i = r^i s$ for all i.

Necessity. If $c_0 = 0$, then $1 \in I_{\mathbf{c}}$, which requires $1x^i \in I_{\mathbf{c}}$, and thus $c_i = 0$ for all i. We choose $s = 0$ and r arbitrary. On the other hand, if $c_0 \neq 0$, then $c_1 - c_0 x \in I_{\mathbf{c}}$, which requires $(c_1 - c_0 x)x^i \in I_{\mathbf{c}}$. We conclude that $c_i c_1 - c_{i+1} c_0 = 0$, and therefore $c_{i+1} = rc_i$, where $r = c_1/c_0$. This yields $c_i = r^i c_0$ for $i \geq 1$. Now r must be an integer, since otherwise c_i will not be an integer when i is sufficiently large. We choose $s = c_0$.

Sufficiency. Suppose that $c_i = r^i s$ for fixed integers r and s. If $c_0 = 0$, then $I_{\mathbf{c}} = \mathbb{Z}[x]$. If $c_0 \neq 0$, then $f \in I_{\mathbf{c}}$ if and only if $f(r) = 0$. Thus $I_{\mathbf{c}}$ is the set of all polynomials in $\mathbb{Z}[x]$ that vanish at r; this set is an ideal of $\mathbb{Z}[x]$.

Solved also by R. Barbara (Lebanon), P. Budney, L. Cagliero & J. Lauret (Argentina), A. E. Caicedo Núñez (Colombia), R. M. Carroll, R. J. Chapman (U. K.), R. Ehrenborg (Canada), D. Faurot, S. M. Gagola Jr., R. Holzsager, U. Klein (Germany), O. P. Lossers (The Netherlands), P. J. Morandi, I. Nemes (Austria), A. Nijenhuis, A. Pedersen (Denmark), R. Stong, A. N. 't Woord (The Netherlands), and the proposer.

A Sequence of Reducible Polynomials

10423 [1993, 1014]. *Proposed by M. Filaseta & C. Nicol, University of South Carolina, Columbia, SC.* For a positive integer n, let

$$P_n(x) = \sum \left\{x^{j-1} : 1 \leq j \leq n, \gcd(j, n) = 1\right\}.$$

For example, $P_1(x) = P_2(x) = 1$, $P_3(x) = x+1$, $P_4(x) = x^2+1$, $P_5(x) = x^3+x^2+x+1$, and $P_6(x) = x^4 + 1$. Prove that $P_n(x)$ is reducible over the rationals for every $n \geq 7$.

Composite solution by Roy Barbara, Lebanese University, Fanar, Lebanon, and Michael Reid, Brown University, Providence, RI. We prove inductively that for $n > 2$, $P_n(x)$ has a factor of the form $1 + x^r$ for some $r > 0$, and it is a nontrivial factor if $n > 6$. If n is prime, then $P_n(x) = \sum_{i=0}^{n-2} x^i = (1+x)(1+x^2+x^4+\cdots+x^{n-3})$, and the factors are nontrivial if $n > 3$. The case $n = 4$ can be checked explicitly.

When $n > 4$ is not prime, we write $n = mp$, where p is prime and $m > 2$. The induction hypothesis yields a factor $1 + x^r$ of $P_m(x)$. Let $A_n = \{j \in \mathbb{Z} : 1 \leq j \leq n, \gcd(j, n) = 1\}$.

If p divides m, then A_n is the disjoint union of translates of A_m by multiples of m, and $P_n(x) = (\sum_{i=0}^{p-1} x^{im}) P_m(x)$. Thus $1 + x^r$ is the desired factor of $P_n(x)$.

If p does not divide m, then within the universe of positive integers less than n,

$$A_n = \{j : \gcd(j, m) = 1\} - \{j : \gcd(j, mp) = p\}.$$

This consists of translates of A_m, less omissions of the form $j = kp$ with $\gcd(k, m) = 1$. Since $x^{kp-1} = x^{p-1}(x^p)^{k-1}$, we obtain

$$P_n(x) = \left(\sum_{i=0}^{p-1} x^{im}\right) P_m(x) - x^{p-1} P_m(x^p). \qquad (*)$$

If p is odd, then $1+x^r$ is a factor of $1+x^{pr}$. Since $1+(x^p)^r$ divides $P_m(x^p)$, equation $(*)$ yields $1 + x^r$ as the desired factor of $P_n(x)$. If n is square-free and has no such odd divisor p, then $n = 2m$, where m is an odd prime. Now explicitly $P_n(x) = (\sum_{i=0}^{m-1} x^{2i}) - x^{m-1} = (1 + x^{m+1})(1 + x^2 + \ldots + x^{m-3})$. The factors are nontrivial if $m > 3$.

Solved also by M. Benedicty, D. Callan, R. J. Chapman (U. K.), N. Komanda, O. P. Lossers (The Netherlands), and the proposers.

Solutions to a Determinantal Equation

10427 [1995, 71]. *Proposed by George Soules, CCR-IDA, Princeton, NJ.* Let A be an n-by-n positive semi-definite Hermitian matrix. Write $A = L + D + L^*$, where L is lower triangular with zero diagonal, and D is the diagonal of A (and L^* is the complex conjugate transpose of L). If $\det(D) \neq 0$, show that all n roots of $\det(zL + zD + L^*) = 0$ lie in the unit disk $|z| \leq 1$. Also, determine when this polynomial can have a root with $|z| = 1$.

Solution by Edward A. Bender, University of California, La Jolla, CA. Let $B = zL+zD+L^*$ and $B^* = L + \bar{z}D + \bar{z}L^*$. Up to a scalar, there is only one linear combination $\alpha B + \beta B^*$ for which the coefficients of L and L^* are equal; the result is

$$(1 - \bar{z})B + (1 - z)B^* = (1 - z\bar{z})A - |1 - z|^2 D. \qquad (*)$$

Given $\det(B) = 0$, choose $v \neq 0$ such that $Bv = 0$. Since $v^*B^* = 0$, the quadratic forms v^*Bv and v^*B^*v are both zero. From equation $(*)$, we obtain

$$(1 - z\bar{z})v^*Av = |1 - z|^2 v^*Dv.$$

Since $a_{ii} > 0$ for every i, the right side is positive for $z \neq 1$. Since x^*Ax is a nonnegative quadratic form, we conclude that $v^*Av > 0$ and $z\bar{z} < 1$. If $z = 1$, then A is singular.

Solved also by O. Krafft & M. Schaefer (Germany), and J. H. Lindsey II.

More Intertwined Exponentials

10475 [1995, 746]. *Proposed by Wu Wei Chao, He Nan Normal University, Xin Xiang City, He Nan Province, China.* For $0 < x < y < 1$ or $1 < x < y$, prove that

$$y^{x^y} \,/\, x^{y^x} > y \,/\, x \qquad (1)$$

and

$$y \,/\, x > y^x \,/\, x^y. \qquad (2)$$

Solution by J. S. Frame, Michigan State University, East Lansing, MI. The assumptions on x and y may be written as $0 < x < y$ and $(x - 1)(y - 1) > 0$.

Taking logarithms, we replace the desired inequalities by

$$x^y \ln y - y^x \ln x > \ln y - \ln x, \qquad (1)^*$$

$$\ln y - \ln x > x \ln y - y \ln x. \qquad (2)^*$$

Since $(x - 1)(y - 1) > 0$, inequalities (2) and (2)* are equivalent to

$$\frac{\ln x}{x - 1} > \frac{\ln y}{y - 1}. \qquad (2)^{**}$$

We prove (2)** by noting that, for $z > 0$ and $z \neq 1$, the positive-valued function $f(z) = (\ln z)/(z - 1)$ has the negative-valued derivative

$$\frac{\ln\left(1 - (1 - z^{-1})\right) + (1 - z^{-1})}{(z - 1)^2},$$

so that $f(x) > f(y)$. This proves (2)**, and hence also (2).

We now introduce a variable t with $0 < t < 1$. From (2), we get $(y/x)^{1/t} > y/x > y^x/x^y$ and $y/x > y^{xt}/x^{yt}$, so that

$$0 < \int_0^1 yx^{yt} - xy^{xt}\, dt = \frac{x^y - 1}{\ln x} - \frac{y^x - 1}{\ln y}, \qquad (1)^{**}$$

which is equivalent to inequalities (1)* and (1) since $\ln x \cdot \ln y > 0$.

Editorial comment. Joe Howard demonstrated how to modify the published solution of MONTHLY Problem E 3291 [1988, 872; 1990, 346] to obtain the slightly stronger inequality (1) of the present problem.

Solved also by R. J. Chapman (U. K.), J. Howard, W. Janous (Austria), P. McCartney, L. Scribani (South Africa), H.-J. Seiffert (Germany), F. Qi (China), and the proposer.

REVIEWS

Edited by **Underwood Dudley**
Mathematics Department, De Pauw University, Greencastle, IN 46135

Emblems of Mind: The Inner Life of Music and Mathematics. By Edward Rothstein. Times Books, 1995, xx + 263, $25.00

Reviewed by **Jeffrey Nunemacher**

That mathematics and classical music share many fundamental attributes is no surprise to anyone acquainted with both fields. Both are highly structural; both can be abstract and unworldly; both depend more on natural talent than on human experience, so both engender prodigies. But is there a connection between mathematics and music that lies deeper than these surface similarities? Just such an exploration is what Edward Rothstein has attempted in his new book *Emblems of Mind*. The result is a speculative rumination on both fields, with illuminating extended examples chosen to illustrate his points. By design, the book is philosophical and metaphorical in nature, so a reader needs to be responsive to poetry, philosophy, and "high art" to be in tune with the author's goal. The title, for example, is drawn from the description of a quest for unifying understanding in Wordsworth's poem "The Prelude", and a romantic, somewhat mystical impulse pervades the book.

At the time of publication of the book, Mr. Rothstein was chief music critic for *The New York Times*; he now writes generally about cultural issues for that distinguished newspaper. Mr. Rothstein holds a Ph.D. from the Committee on Social Thought at the University of Chicago, and this book, which attempts to relate his two enthusiasms, derives (at least in spirit) from his study there. He has studied mathematics and music at Yale, Brandeis, and Columbia and has a good understanding of the nature of both fields. On the other hand, he has never been a professional mathematician, so his views of this subject, though they are well-informed and rewarding to ponder, differ somewhat in substance and definitely in tone from those held by many mathematicians. In particular, the author's approach is heavily philosophical, and I suspect that many mathematical readers maintain an impatience and skepticism towards the diffuse goals and perceived "woolliness" of philosophy.

The book is organized into six chapters. The first and the last (entitled "Prelude" and "Chorale") introduce and recapitulate his main themes. The second ("Partita") explores topics in mathematics with a view to musical analogies; the third ("Sonata") does the same for music with mathematical analogies. The fourth ("Theme and Variations") and fifth ("Fugue") address the issues of beauty and truth in both areas. Though the author claims that he writes for a literate general reader, the book is most likely to appeal to someone who has had some serious involvement with both fields. I think it is unlikely that a reader who has never studied music theory would find much of what is said about music very compelling. There is also a progression of ideas in the book from the concrete and definite to

the abstract and mystical. Mr. Rothstein finds the real essence and inner life of both mathematics and music to be the reflection of Platonic "Forms", which he discusses in the final chapter.

The mathematics treated in the book will generally not be new to a mathematically trained reader. There is some discussion of the axiomatic method and its limitations, of some problems in elementary number theory, of topology and the problem of homeomorphism, and of both the standard and nonstandard models of the continuum. These expositions are nicely done and the general reader can catch the flavor of the subject. The author does not make the mistake of inferring from the formality of most mathematical exposition that what mathematicians really do is to play with axiom systems. It is true, however, that the author's interests are in pure mathematics, which does bias his point of view. He views mathematics as a humanity rather than a science; thus the interesting issues are of style, taste, and history rather than applicability. I wonder, for instance, whether a physicist or applied mathematician would agree with his statement that "truth in mathematics is often quite different from truth in physics". There are nice insights about issues not usually discussed by mathematicians. For example, in a discussion of the Chinese Remainder Theorem, drawn from Davis and Hersh [1], Mr. Rothstein gives four quite different formulations appropriate to different historical eras and to varying levels of abstraction. He then analyzes the issue of style in mathematics and its similarity to musical styles.

The analysis of beauty in Chapter 4 is devoted mostly to making the case for the significance of beauty within mathematics, since most readers will grant that beauty is a driving force behind music. Mr. Rothstein finds beauty in a variety of guises and purposes: the esthetics of discovery (Poincaré), of written exposition (Maxwell), or oral exposition (his teachers at Yale—Robinson and Kakutani), of mathematical argument (Péter and Gauss). This range of examples is fruitful to consider and sheds new light on the issue. But when the author moves from specific instances to general assertions about his topic, one finds poetic assertions such as "The beauty of the true is an inescapable aspect of the inner life of mathematics and music" (p. 146). Such statements verge on mysticism rather than useful substantive analysis. It is hard to find a definition of truth that fits both mathematics and music and is not so vague as to be meaningless.

There are a few errors in the book, one of which is surprising given the author's general level of mathematical competence. On page 71 the epsilon-delta definition of continuity is stated incorrectly (with the most common error that an undergraduate analysis student would make). On page 75 the impression is given that there are two nondegenerate classes of loops on a torus rather than the infinite number that lie in the fundamental group. Overall, though, a mathematician will be happy with the accuracy and tone of the writing about mathematics.

The book was successful in getting this reader back to his piano to study Bach's Fugue in D♯ Minor from the Well-Tempered Clavichord, Book 1, which is much discussed in Chapter 3 as an example of melody and counterpoint. Actually, it took some effort to locate this piece. Eventually I found the composition as Fugue No. VIII in the Schirmer (Czerny) edition but with a key signature of E♭ Minor. What can explain the change of key from D♯? In the authoritative Bach-Gesellschaft edition, the associated prelude is notated in E♭ with the fugue in D♯; one presumes that later editions have amended the notation for consistency. The author's comments about the fugue are much more illuminating in front of a piano. Another composition that arises in the text and is considerably easier to play for readers who have studied some piano is Chopin's Prelude in A Minor. This piece is

short enough to be reproduced on one page of Mr. Rothstein's book and is explored as an example of disturbing and dissonant beauty.

How successful is Mr. Rothstein's attempt to find deep common ground between mathematics and music? He looks in promising places, such as the unifying concepts of mapping and transformations, which occur in both fields, but he does not find much apart from analogy. Ultimately the author locates the object of his quest in Platonic philosophy, which is sufficiently indefinite to be unsatisfying. I did not become as absorbed in this book as I had expected. While I found individual sections of the book insightful and respect the author's attempt to articulate the common underlying principles in the two fields, I am not persuaded that there is much here beyond metaphor. For catching the spirit of mathematics and examining it from a humanistic perspective, the book by Davis and Hersh is more successful. Perhaps readers who are less suspicious of metaphor as an analytic device will think differently. I urge you to sample this book for yourself.

REFERENCE

1. Philip Davis and Reuben Hersh. *The Mathematical Experience*. Birkhäuser, Boston, 1981.

Department of Mathematical Sciences
Ohio Wesleyan University
Delaware, Ohio 43015
jlnunema@cc.owu.edu

A Tour of the Calculus. By David Berlinski. Pantheon Books, 1995, xvii + 332, $27.50

Reviewed by **Israel Kleiner**

Yet another calculus book? Yes and no.

If you are looking for a book in which the formulas, equations, theorems, and proofs take center stage, this book will add little. If you want a tour of the subject that offers a wider perspective, Berlinski invites you to come aboard.

Here is the author's nutshell view of calculus, "the massive load-bearing walls and buttresses of the subject":

> The overall structure of the calculus is simple. The subject is defined by *a fantastic leading idea*, [namely that] the real world may be understood in terms of the real numbers; *one basic axiom*, [which] brings the real numbers into existence; *a calm and profound intellectual invention*: the mathematical function; *a deep property*: continuity; *two central definitions*: instantaneous speed and the area underneath a curve; *one ancillary definition*: limit; *one major theorem*: the mean value theorem; and *the fundamental theorem of the calculus*.

The twenty-six chapters of this 300-page book flesh out this skeleton in an engaging, witty, and mathematically honest manner.

Berlinski's outline provides the contours for the *logical* development of calculus rather than for its *historical* evolution. For example, the real numbers came *last* in the historical sequence, and the notion of function was not available to the creators of calculus! Berlinski is aware of this: "In its historical development, the

calculus represents an exercise in delayed gratification", he affirms in his inimitable way. But his book is *not* an historical account of calculus.

The history of the subject is, however, given consideration in *A Tour of the Calculus*. As it must (it seems to me). For the author deals with central *ideas* of the subject, and history supplies the context and motivation for their rise and flowering. It also points to the subject as a human enterprise, standing on the shoulders of giants. The author's pronouncements on some of these giants are variously endearing ("In his death as in his life, Descartes was unique: he was the only great thinker to die of discomfort"), intriguing ("[Euler] amused himself with mathematical oddities, his intelligence functioning in that strange frictionless world which in the history of mathematics is inhabited only by Euler himself and in the history of music only by Mozart"), and insightful ("[Riemann] was in his temperament a geometer, in his affiliations a Platonist, and in his soul a visionary").

Practically all of the prominent mathematicians of Europe around 1650 could solve many of the problems in which elementary calculus is now used. Why, then, is the subject considered to have been invented (by Newton and Leibniz) in the *last third* of that century? This question helps to focus on some of the central ideas of calculus, for example the recognition that tangents and velocities on the one hand, and areas and arc-lengths on the other, can be subsumed under general concepts: the derivative and the integral, respectively. Following the subject's invention (discovery?), it took another two centuries to provide it with rigorous foundations. "Certain intellectual tools may be successfully used before they are successfully understood" is how the author puts it. Cause for reflection.

Among other ideas that Berlinski reflects on are the roles of arithmetic versus geometry, the nature of $\sqrt{2}$, the purpose of symbols, and the significance of the notions of function, continuity, derivative, and integral. He lays special emphasis on continuity—of the line and of a function:

> Continuity is an aspect of things as rooted in reality as the fact that material objects occupy space; it is the contrast between the continuous and the discrete that is the great generating engine by which the real numbers are constructed and the calculus created. The concept of continuity is, like so many profound concepts, both simple and elusive, elementary and divinely enigmatic.

In the 17th century Leibniz enunciated an all-embracing Principle of Continuity, which, in a mathematical context, said roughly that what holds in a given case also holds in what appear to be like cases. At one point Leibniz justified the rules of operation with infinitesimals by their resemblance to those with real numbers. The principle played a significant role in 18th- and 19th-century mathematics. In 18th-century calculus the modus operandi was that what is true of the finite (e.g., polynomials) is true also of the infinite (e.g., power series). In early 19th-century algebra the principle of continuity went under the name of the Principle of Permanence of Equivalent Forms. It said, essentially, that the laws of operation with positive numbers carry over to negative numbers. Poncelet in the early 19th century proclaimed a principle of continuity in projective geometry. The basic idea in all of these was that what were viewed to be insignificant changes in input had no effect on output.

The concept of continuity of a *function*, defined in terms of epsilons and deltas, came into being in the second half of the 19th century—about 200 years following the invention of calculus. Cauchy did define continuity in 1821, but in terms of infinitesimals ("an infinitely small increment of the variable always produces an infinitely small increment of the function"). This got him into trouble: he "proved"

(for example) that a convergent series of continuous functions is continuous. (To be sure, some have argued that it is a false reading of history to view Cauchy's proof as erroneous.) Mathematicians believed, and some "proved", that continuity implies differentiability (except possibly at a finite number of points)—until Riemann and Weierstrass jolted the mathematical community with their examples of continuous *nowhere* differentiable functions. And continuity was sometimes identified with the Intermediate Value Property (a counterexample was given by Darboux in 1870). "Elusive" and "enigmatic", indeed, that continuity.

But not just continuity. Even functionality—nowadays elementary and simple and commonplace—proved to be elusive and controversial, "purely an intellectual object... and one of the great imponderables that like certain movie stars is forever familiar but forever unknown".

Recorded mathematical history goes back almost 4000 years. During the first 3700 of these, mathematicians developed the elements of algebra, deductive geometry, trigonometry, analytic geometry, and calculus. Surprisingly, the notion of function was not part of that development. The concept originated in the early 18th century, well into the so-called modern period in the evolution of mathematics. Why so late? Principally because there was no manifest need for it earlier. (Is there a lesson here for pedagogy?) And secondarily because the algebraic prerequisites were lacking—the coming to terms with the continuum of real numbers, and the development of symbolic notation.

When the function concept did emerge, it was first as a "formula" (a so-called "analytic expression", although what that meant was not made precise), then as a curve, a formula again, an arbitrary correspondence, a formula, a set of ordered pairs, a generalized function .... The reception accorded the introduction of new functions was not always cordial. For example, Hermite called continuous, nowhere-differentiable functions "a lamentable evil", and Poincaré termed them "monsters". Such an elementary notion, that of function, yet such a rich and tortuous history.

Hilbert noted that every mathematical theory goes through three stages of development—the naïve, the formal, and the critical. The same can be said of the evolution of mathematical concepts, for example the derivative: from tangent and velocity (naïve), through differential and fluxion (formal), to limit of a difference quotient (critical). The psychological (hence pedagogical?) distance between the naïve and critical stages is forbidding, according to the author:

> The derivative is an artifact, the first of the great concepts of modern science that fails conspicuously to correspond to anything in real life. In order to express speed as a function of time, the mathematician is prepared to sacrifice common sense, he is prepared to sacrifice the intuitive definition of speed, in plain fact, he is prepared to sacrifice everything.

Let us set aside discussion of other issues raised in Berlinski's book and ask: What is the book good for? And whom is it good for? Here is the author's answer:

> I have written this book for men and women who wish to understand the calculus as an achievement in human thought. It will not make them mathematicians, but I suspect that what they want is simply a little more light shed on a dark subject.

Calculus as an achievement in human thought? That as a major goal would seem to disqualify the book as a text for most (all?) existing calculus courses. Yes, calculus is a set of rules or algorithms (a "calculus"); it is also a theory to explain why the rules work; and it is applications (of the theory and the rules) to fundamental problems in science. But calculus *is* one of the great intellectual

accomplishments of civilization. Ought we perhaps not begin every calculus course with some such statement? Of course it is rather difficult to deliver on this promise, but Berlinski's book might help us to start thinking about it.

A Tour of the Calculus is written in narrative style and is easy reading (for a mathematician, that is). There *are* (in a dozen brief appendices attached to relevant chapters) formal definitions of major concepts, and proofs of a number of important results. But this is no systematic account even of elementary calculus. That, too, would seem to disqualify it as a text. We would all agree, however, that among the multitude of definitions and theorems taught in a calculus course some are more important than others. We ought to bear this insight in mind and pass it on to students. Berlinski's book might help us focus our thoughts on the issue.

A Tour of the Calculus is one in a genre of recently published (or reprinted) "popular" books. Among my favorites are E. T. Bell, *Mathematics: Queen and Servant of Science*, R. Courant and H. Robbins, *What is Mathematics?*, T. Dantzig, *Number: The Language of Science*, E. Kasner and J. R. Newman, *Mathematics and the Imagination*, V. M. Tikhomirov, *Stories about Maxima and Minima*, O. Toeplitz, *Calculus: A Genetic Approach*, and N. Y. Vilenkin, *In Search of Infinity*. Some of my criteria for selecting them are: they focus on mathematical *ideas* without giving short shrift to the mathematics. They incorporate at least some history of the subject with which they deal (and occasionally a touch of philosophy). They are well written and can be heartily recommended to students. And they make *us*—teachers—think.

I know I am setting myself up for severe reprimand for (a) omitting many deserving books, and (b) including Bell. I hope I can count on the *Monthly* editors to come to my defense for transgression (a), but I'd like to offer a defense of (b).

My colleagues in the history of mathematics especially will take me to task for including Bell. His books are reputed to contain historical inaccuracies (if not worse), and I will not deny that. It has been said unkindly (and I imagine unfairly) that Bell would not let the truth stand in the way of a good story. He *does* write beautifully and inspiringly (*my* opinion, it goes without saying). His *Men* (ouch) *of Mathematics* and his *Development of Mathematics* (groan, fellow historians) have inspired *me* (a good many years ago) to take a serious interest in the history of mathematics, and for that I am eternally grateful to him. With appropriate forewarning, I would not hesitate to recommend Bell's books to aspiring students of the history of mathematics, if only to awaken their interest in the subject.

But I must in conclusion return to Berlinski's book. There is, in general, more than one way of achieving a set of objectives. I imagine that the book's readers may have their own ideas about how to achieve those stated and pursued by the author of *A Tour of the Calculus*, or they may set other goals for such a book. I believe, however, that a reviewer's function is to judge the book that the author has actually written, not one that he or she feels the author *should* have written. In that spirit, I am happy to commend *A Tour of the Calculus* to students and colleagues (especially the latter): it is for the most part pleasurable reading and it may set us to thinking (again) about some of the issues in the learning and teaching of calculus (with a bow to the various calculus-reform movements).

Department of Mathematics and Statistics
York University
North York, Ontario M3J 1P3
Canada
kleiner@yorku.ca

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of
Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, P, L*. *The Parsimonious Universe: Shape and Form in the Natural World.* Stefan Hildebrandt, Anthony Tromba. Springer-Verlag, 1996, xiii + 330 pp, $32.95. [ISBN 0-387-97991-3] A wonderful, visually rich exploration of the worldly consequences of optimization principles (networks, geodesics, surfaces, living forms). Lavishly illustrated with historical, scientific, and computer-generated images. An expanded version of the authors' 1984 *Mathematics and Optimal Form* (TR, August–September 1985; Extended Review, June–July 1988). LAS

Reference, P, L*. *Encyclopedia of Operations Research and Management Science.* Eds: Saul I. Gass, Carl M. Harris. Kluwer Academic, 1996, xxxx + 753 pp, $350. [ISBN 0-7923-9590-5] 185 expository articles plus descriptions, definitions, abbreviations, etc. Articles are designed to serve as initial sources of information. Each includes some background or history, describes applications, and lists references.

Recreational Mathematics, S(13). *Challenging Problems in Geometry.* Alfred S. Posamentier, Charles T. Salkind. Dover, 1996, ix + 245 pp, $7.95 (P). [ISBN 0-486-69154-3] Nearly 200 nonroutine problems in Euclidean geometry. Republication of the *Second Edition* published by Dale Seymour Publications in 1988.

Recreational Mathematics, S(13). *Challenging Problems in Algebra.* Alfred S. Posamentier, Charles T. Salkind. Dover, 1996, x + 262 pp, $8.95 (P). [ISBN 0-486-69148-9] More than 300 problems that require little more than knowledge of high school algebra (and some ingenuity!). Republication of the *Second Edition* published by Dale Seymour Publications in 1988.

Education, P. *Teaching Mathematics: Toward a Sound Alternative.* Brent Davis. Garland Pub, 1996, xxxii + 324 pp, $21.95 (P). [ISBN 0-8153-2298-4] A densely end-noted philosophical apologia for a new approach to mathematics education centered on a theoretical framework the author calls "enactivism"—the idea that understanding flows not from reason or observation but from action. This analysis leads to a pedagogy of active listening from which reason emerges. LAS

Education, P, L. *Communication in Mathematics, K–12 and Beyond.* Portia C. Elliott, Margaret J. Kenney. 1996 Yearbook. NCTM, 1996, x + 256 pp, $22. [ISBN 0-87353-423-9] 28 diverse papers by 55 authors on mathematical discourse, writing, and language all aimed at supporting NCTM's "communication" standard. Zal Usiskin's concluding chapter summarizes much of the volume by describing mathematical language in its various possible forms: written, oral, pictorial, foreign, dead, nonsense, abstract, and native. LAS

Education, P. *Research in Collegiate Mathematics Education. II.* Eds: Jim Kaput, Alan H. Schoenfeld, Ed Dubinsky. CBMS Issues in Math. Educ., V. 6. AMS, 1996, xi + 217 pp, $37 (P). [ISBN 0-8218-0382-4] 10 research papers on diverse topics, and a list of questions for future work.

History, P, L. *The Collected Papers of Albert Einstein, Volume 4, The Swiss Years: Writings, 1912–1914.* Transl: Anna Beck; Consultant: Don Howard. Princeton Univ Pr, 1996, xi + 314 pp, $39.50 (P). [ISBN 0-691-02610-6]

History, P. *Collected Papers, Volume I.* Florentin Smarandache. Editura Tempus Romania, 1995, 301 pp, (P). [ISBN 973-9205-02-X]

Combinatorics, P*, L.** *Handbook of Combinatorics.* Eds: R.L. Graham, M. Grötschel, L. Lovász. MIT Pr, 1995, $300 set [0-262-07169-X]. *Volume 1*, cii + 1018 pp, $175, [ISBN 0-262-07170-3]; *Volume 2*, cii + 1177 pp, $175. [ISBN 0-262-07171-1] A comprehensive overview of the present state of combinatorics. Organized into five sections: Structures, Aspects, Methods, Applications, and Horizons.

Discrete Mathematics, L. *Problems and Exercises in Discrete Mathematics.* G.P. Gavrilov, A.A. Sapozhenko. Texts in Math. Sci., V. 14. Kluwer Academic, 1996, xi + 422 pp, $198. [ISBN 0-7923-4036-1] Problems with hints and solutions on topics such as Boolean algebra, graphs, networks, coding theory, algorithm theory, combinatorics, and logical design. Takes a functional approach to discrete mathematics (typical of the Moscow school). LC

Discrete Mathematics, T(13: 1). *Introductory Discrete Mathematics.* V.K. Balakrishnan. Dover, 1996, xiv + 236 pp, $9.95 (P). [ISBN 0-486-69115-2] Republication, with corrections, of the 1991 Prentice Hall edition. Topics include basic counting principles, permutations, combinations, the inclusion-exclusion principle, generating functions, algorithm analysis, and graph theory.

Number Theory, S, P, L.** *The New Book of Prime Number Records.* Paulo Ribenboim. Springer-Verlag, 1996, xxiv + 541 pp, $59.95. [ISBN 0-387-94457-5] A delightful collection of just about everything dealing with prime numbers. This edition has updated records and a few new sections. (*Second Edition*, TR, February 1990.) CEC

Group Theory, P. *Representations of Infinite-Dimensional Groups.* R.S. Ismagilov. Transl. of Math. Mono., V. 152. AMS, 1996, x + 197 pp, $85. [ISBN 0-8218-0418-9]

Algebra, T(14–16: 1, 2).** *Abstract Algebra: An Introduction, Second Edition.* Thomas W. Hungerford. Saunders College, 1997, xix + 588 pp, $54. [ISBN 0-03-010559-5] Minor organizational changes. Keeps the distinctive organizational structure: integers, then polynomials, then groups and rings. (*First Edition*, TR, May 1990.) TH

Algebra, T*(17–18: 1), P, L. *Field and Galois Theory.* Patrick Morandi. Grad. Texts in Math., V. 167. Springer-Verlag, 1996, xvi + 281 pp, $42.50. [ISBN 0-387-94753-1] A splendid book. Covers separability, Galois theory, applications, and infinite extensions. Large collection of examples and exercises. TH

Algebra, T(18: 1), P. *An Algebraic Introduction to Complex Projective Geometry: 1. Commutative Algebra.* Christian Peskine. Stud. in Adv. Math., V. 47. Cambridge Univ Pr, 1996, x + 230 pp, $39.95. [ISBN 0-521-48072-8] The first of a three-part introduction to commutative algebra developed as a tool for geometry and number theory. Topics include: Noetherian rings and modules, polynomial rings in several variables, Weil and Cartier divisors. TH

Algebra, P. *Period Spaces for p-divisible Groups.* M. Rapoport, Th. Zink. Annals of Math. Stud., No. 141. Princeton Univ Pr, 1996, xxi + 324 pp, $59.50 (P). [ISBN 0-691-02781-1]

Complex Analysis, P. *Lectures on Entire Functions.* B. Ya. Levin. Transl. of Math. Mono., V. 150. AMS, 1996, xv + 248 pp, $99. [ISBN 0-8218-0282-8]

Differential Equations, T(14: 1). *A First Course in Differential Equations with Modeling Applications, Sixth Edition.* Dennis G. Zill. Brooks/Cole, 1997, xiii + 438 pp, $67.95. [ISBN 0-534-95574-6] Increased emphasis of modeling; color plates with new applications; more emphasis on non-linear and systems of differential equations. (*Fourth Edition*, TR, February 1989.) LC

Differential Equations, P. *Asymptotic Solutions of the One-Dimensional Schrödinger Equation.* S. Yu. Slavyanov. Transl. of Math. Mono., V. 151. AMS, 1996, xvi + 190 pp, $99. [ISBN 0-8218-0563-3]

Differential Equations, P. *Linear Differential Operators.* Cornelius Lanczos. Classics in Appl. Math., V. 18. SIAM, 1996, xvii + 564 pp, $49.50 (P). [ISBN 0-89871-370-6]

Dynamical Systems, P. *Introduction to the Qualitative Theory of Dynamical Systems on Surfaces.* S. Kh. Aranson, G.R. Belitsky, E.V. Zhuzhoma. Transl. of Math. Mono., V. 153. AMS, 1996, xiii + 325 pp, $129. [ISBN 0-8218-0369-7]

Functional Analysis, P. *Index Theory, Coarse Geometry, and Topology of Manifolds.* John Roe. CBMS Reg. Conf. Ser. in Math., No. 90. AMS, 1996, ix + 100 pp, $17 (P). [ISBN 0-8218-0413-8]

Functional Analysis, P. *C_0-Groups, Commutator Methods and Spectral Theory of N-Body*

Hamiltonians. Werner O. Amrein, Anne Boutet de Monvel, Vladimir Georgescu. Progress in Math., V. 135. Birkhäuser Boston, 1996, xiv + 460 pp, $98. [ISBN 0-8176-5365-1]

Functional Analysis, P. *Partial Differential Equations and Functional Analysis: In Memory of Pierre Grisvard.* Eds: Jean Cea, *et al.* Progress in Nonlinear Diff. Eqns. & Their Applic., V. 22. Birkhäuser Boston, 1996, xxii + 263 pp, $134.50. [ISBN 0-8176-3839-3] Papers from a 1994 conference together with a bibliography of Grisvard's works and one of his previously unpublished papers.

Analysis, P. *Trigonometric Fourier Series and Their Conjugates.* Levan Zhizhiashvili. Math. & Its Applic., V. 372. Kluwer Academic, 1996, xii + 300 pp, $165. [ISBN 0-7923-4088-4]

Differential Geometry, P. *Finsler Geometry.* Eds: David Bao, Shiing-shen Chern, Zhongmin Shen. Contemp. Math., V. 196. AMS, 1996, xxiii + 310 pp, $61 (P). [ISBN 0-8218-0507-X] Proceedings of a 1995 AMS–IMS–SIAM Joint Summer Research Conference at the University of Washington.

Differential Geometry, P. *Metrics, Connections and Gluing Theorems.* Clifford Henry Taubes. CBMS Reg. Conf. Ser. in Math., No. 89. AMS, 1996, v + 90 pp, $15 (P). [ISBN 0-8218-0323-9]

Differential Geometry, P. *Riemannian Geometry.* Takashi Sakai. Transl: Takashi Sakai. Transl. of Math. Mono., V. 149. AMS, 1996, xiii + 358 pp, $119. [ISBN 0-8218-0284-4]

Differential Geometry, P. *Some Questions of Differential Geometry in the Large.* Ed: E.V. Shikin. AMS Transl., Ser. 2, V. 176. AMS, 1996, x + 192 pp, $89. [ISBN 0-8218-7506-X] 6 papers present recent results from Russia and the Ukraine.

Geometry, P, L. *Strange Phenomena in Convex and Discrete Geometry.* Chuanming Zong. Ed: James J. Dudziak. Universitext. Springer-Verlag, 1996, x + 158 pp, $29 (P). [ISBN 0-387-94734-5] Famous problems in convex and discrete geometry that have counterintuitive and/or strange answers. Contains recent advances and an excellent bibliography. CEC

Algebraic Topology, T(16–17: 2). *Algebraic Topology.* C.R.F. Maunder. Dover, 1996, vii + 375 pp, $10.95 (P). [ISBN 0-486-69131-4] Republication of the 1980 Cambridge University Press corrected printing. (1970 Van Nostrand Reinhold edition, TR, February 1972; Extended Review, April 1973.)

Topology, P. *Renormalization and 3-Manifolds which Fiber over the Circle.* Curtis T. McMullen. Annals of Math. Stud., No. 142. Princeton Univ Pr, 1996, vii + 253 pp, $24.95 (P); $55. [ISBN 0-691-01153-2; 0-691-01154-0]

Mathematical Modeling, T?(16–17: 1). *Elements of Pattern Theory.* Ulf Grenander. Johns Hopkins Univ Pr, 1996, xiii + 222 pp, $24.95 (P); $65. [ISBN 0-8018-5188-2; 0-8018-5187-4] Not about pattern recognition algorithms. Uses algebraic and probabilistic methods to "formalize the ... concept of pattern in terms of a mathematical framework." Some background in algebra, probability theory, and some programming experience needed. LB

Stochastic Processes, P. *Schrödinger Diffusion Processes.* Robert Aebi. Prob. & Its Applic. Birkhäuser Boston, 1996, viii + 186 pp, $79.50. [ISBN 0-8176-5386-4]

Elementary Statistics, T(13–14: 1, 2). *Statistics: Basic Principles and Applications.* William J. Adams, Irwin Kabus, Mitchell P. Preiss. Kendall/Hunt, 1994, xv + 792 pp, $60.95, [ISBN 0-8403-8964-7]; *Companion to Statistics: Basic Principles and Applications,* 1994, vii + 215 pp, $18.95 (P). [ISBN 0-8403-9414-4] Non-calculus-based. Covers basic descriptive and inferential statistics, with additional sections on index numbers, time series, nonparametrics. ANOVA, contingency tables, goodness-of-fit covered briefly in a chapter on additional hypothesis tests. Uses hypothetical examples rather than real-world data. LB

Statistical Methods, P. *Lecture Notes in Control and Information Sciences–216: Recursive Nonlinear Estimation: A Geometric Approach.* Rudolf Kulhavý. Springer-Verlag, 1996, xvi + 224 pp, $54 (P). [ISBN 3-540-76063-6]

Statistics, P. *Mathematical Theory of Reliability.* Richard E. Barlow, Frank Proschan. Classics in Appl. Math., V. 17. SIAM, 1996, xv + 258 pp, $34.50 (P). [ISBN 0-89871-369-2]

Computer Systems, P. *Java Class Reference Package.* Randy Chapman. Specialized Systems Consultants, 1996, $7 (P) set, [ISBN 0-916151-96-4]. *applet, awt and util Class Reference,* 20 pp, (P); *lang, io and net Class Reference,* 18 pp, (P).

Computer Science, C, P, L. *Encyclopedia of Graphics File Formats, Second Edition.* James D. Murray, William vanRyper. O'Reilly & Associates, 1996, xxxvi + 1116 pp, $79.95 (P), with CD ROM. [ISBN 1-56592-161-5]

Computer Science, P. *Statistical Software Engineering.* National Academy Pr, 1996, x + 73 pp, $29 (P). [ISBN 0-309-05344-7] A re-

port on opportunities for statistical thinking to contribute to software engineering.

Computer Science, P. *Comparative Concurrency Semantics and Refinement of Actions.* R.J.H. van Glabbeek. CWI Tract, V. 109. Stichting Mathematisch Centrum, 1996, 285 pp, Dfl. 50 (P). [ISBN 90-6196-454-7]

Computer Science, P. *Probabilistic Expert Systems.* Glenn Shafer. CBMS–NSF Reg. Conf. Ser. in Appl. Math., V. 67. SIAM, 1996, viii + 80 pp, $24.50 (P). [ISBN 0-89871-373-0]

Computer Science, P. *Network and Internet Security.* Vijay Ahuja. Academic Pr, 1996, xix + 324 pp, $219.95 (P). [ISBN 0-12-045595-1]

Computer Science, P. *Getting Connected: The Internet at 56K and Up.* Kevin Dowd. O'Reilly & Associates, 1996, xii + 410 pp, $29.95 (P). [ISBN 1-56592-154-2]

Applications (Biological Science), P. *Mathematics and Physics of Emerging Biomedical Imaging.* National Academy Pr, 1996, xvii + 238 pp, $29 (P). [ISBN 0-309-05387-0] Surveys contributions of the mathematical sciences and physics to the current state of dynamic biomedical imaging and outlines research opportunities.

Applications (Biological Science), T(14–15: 1), L. *An Introduction to the Mathematics of Biology: With Computer Algebra Models.* Edward K. Yeargers, Ronald W. Shonkwiler, James V. Herod. Birkhäuser Boston, 1996, x + 417 pp, $64.50. [ISBN 0-8176-3809-1] Mathematics behind problems such as aging, genetics, HIV, neurophysiology, etc. Biological concepts developed as needed. Maple code interspersed. Rich source of examples for teachers. Prerequisite: one year of calculus, some linear algebra; statistics introduced in text. LC

Applications (Economics), T(17–18: 1), P. *Modeling and Optimization of the Lifetime of Technologies.* Natali Hritonenko, Yuri Yatsenko. Appl. Optim., V. 4. Kluwer Academic, 1996, xxxvi + 249 pp, $130. [ISBN 0-7923-4014-0] Application of modeling, optimization, systems analysis, and control theory to economics of replacement and renewal of technologies in industry. LB

Applications (Physics), P. *Large-Scale Structures in Acoustics and Electromagnetics.* National Academy Pr, 1996, x + 252 pp, $29 (P). [ISBN 0-309-05337-4] Proceedings of a 1994 symposium. Papers focus on computational methods for determining the dynamics of large-scale systems.

Applications (Quantum Theory), P. *Contemporary Mathematical Physics: F.A. Berezin Memorial Volume.* Eds: R.L. Dobrushin, *et al.* AMS Transl. Ser. 2, V. 175. AMS, 1996, ix + 236 pp, $99. [ISBN 0-8218-0426-X] 12 papers in group representation theory, supermathematics, and spectral analysis. Also a brief scientific biography of Berezin and some personal recollections.

Applications (Quantum Theory), P. *Quantization, Nonlinear Partial Differential Equations, and Operator Algebra.* Eds: William Arveson, Thomas Branson, Irving Segal. Proc. of Symp. in Pure Math., V. 59. AMS, 1996, x + 224 pp, $54. [ISBN 0-8218-0381-6] Proceedings of the 1994 John von Neumann Symposium at MIT.

Applications (Systems Theory), S(15–16), P, L*. *Would-Be Worlds: How Simulation Is Changing the Frontiers of Science.* John L. Casti. Wiley, 1997, xii + 242 pp, $24.95. [ISBN 0-471-12308-0] A preview of the 21st-century science of complex systems, which is "still awaiting its Newton." For lack of a theory, computer simulation substitutes. Such systems—physical, biological, behavioral, social—feature a modest number of intelligent, adaptive agents that react to local (rather than global) information. Derived from work at the Santa Fe Institute; illustrated with diverse models ranging from football to linguistics, from traffic flow to neural networks. LAS

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–215: Colloquium on Automatic Control.* Eds: Claudio Bonivento, Giovanni Marro, Roberto Zanasi. Springer-Verlag, 1996, x + 226 pp, $54 (P). [ISBN 3-540-76060-1] Invited papers from a 1996 event at the University of Bologna.

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–218: L_2-Gain and Passivity Techniques in Nonlinear Control.* Arjan van der Schaft. Springer-Verlag, 1996, 168 pp, $43 (P). [ISBN 3-540-76074-1]

Applications (Systems Theory), P. *Sampling in Digital Signal Processing and Control.* Arie Feuer, Graham C. Goodwin. Systems & Control: Found. & Applic. Birkhäuser Boston, 1996, xxxii + 541 pp, $74.50. [ISBN 0-8176-3934-9]

Reviewers

LB: Lynne Baur, Carleton; LC: Laura Chihara, St. Olaf; CEC: Clifton E. Corzatt, St. Olaf; TH: Tom Halverson, Macalester; LAS: Lynn Arthur Steen, St. Olaf.

THE AUTHORS

JIM PITMAN acquired an appreciation of combinatorics and probability from his father Edwin J. G. Pitman, in Hobart, Tasmania. He was an undergraduate at the Australian National University in Canberra, and received a Ph.D. from the University of Sheffield. After spending two years in Cambridge he settled in Berkeley, California. His research involves the asymptotic description of random combinatorial objects, such as partitions, permutations, trees and mappings, in terms of Brownian motion and Poisson processes.

KAREN HUNGER PARSHALL is Associate Professor of Mathematics and History at the University of Virginia and editor of *Historia Mathematica*. Her research focuses primarily on the history of nineteenth- and early twentieth-century algebra. She is the co-author (with David E. Rowe) of *The Emergence of the American Mathematical Research Community, 1876–1900*, has recently completed an edition with historical and mathematical commentary of the selected correspondence of J. J. Sylvester, and is now at work on a full-scale biography of Sylvester.

EUGENE SENETA is Professor in the Department of Mathematics and Statistics at the University of Sydney and a fellow of the Australian Academy of Science. His research interests are in branching processes, nonnegative matrices, regularly varying functions, and the history of probability and statistics, especially in France and Eastern Europe. He is the co-author (with C. C. Heyde) of *I. J. Bienaymé: Statistical Theory Anticipated.*

GEOFFREY GOODSON was educated in Great Britain, obtaining a B.Sc. degree at Hull University, an M.Sc. at Warwick University, and a D.Phil. at the University of Sussex. From 1972 until 1989, he taught in South Africa at the Universities of Cape Town and Witwatersrand. He is now teaching at Towson State University. His academic interests include ergodic theory, dynamical systems, and operator theory. His main non-academic interests are his family and playing squash. He was Maryland 40^+ squash champion in 1992.

THOMAS TUCKER received his BA from Harvard University in 1967 and his PhD from Dartmouth College in 1971. He has been at Colgate University since 1973, where he is the Charles G. Hetherington Professor of Mathematics. His research interests are low-dimensional topology and topological graph theory. For the MAA, he has served as Vice President, chaired the committee on Calculus Reform and the First Two Years (CRAFTY), and edited *Priming the Calculus Pump*. He is a co-author of the textbook developed by the Calculus Consortium based at Harvard. Although his real ambition was to be a TV weatherperson, he ended up in mathematics just like his grandfather, father, brother, and son.

HOWARD SWANN obtained his Ph.D. in applied mathematics from the University of California at Berkeley. He is co-author of *Prof. E. McSquared's Original, Fantastic & Highly Edifying Calculus Primer*, now out in its new *Expanded Intergalactic Version*. His research is concerned with contriving new computer algorithms for approximating solutions to partial differential equations.

EDWARD B. BURGER received his Ph.D. from the University of Texas at Austin in 1990. He was a postdoctoral Fellow at the University of Waterloo before arriving at Williams College. His research interests are in number theory. At the 1996 MathFest he co-produced and co-starred in the first ever math comedy play (see reviews in the Nov. 1996 *Notices of the AMS* and in Ivars Peterson's MathLand column at the MAA web site at http://www.maa.org/.) He is the co-author of two books: *The Heart of Mathematics: An invitation to effective thinking* and *The Invisible Art*, both to be published in 1998.

FRANK MORGAN works in minimal surfaces. He has a weekly live call-in Math Chat on local cable TV, featured in Ivars Peterson's column MathLand at the MAA web site at http://www.maa.org/, and a biweekly Math Chat column in The Christian Science Monitor, sometimes available via the web page http://www.csmonitor.com/. His books include *Geometric Measure Theory*: *a Beginner's Guide*, *Riemannian Geometry*: *a Beginner's Guide*, and *Calculus Lite*.

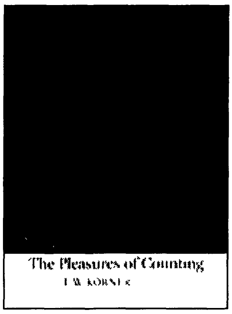
JEFFREY NUNEMACHER entered Oberlin Conservatory as an organ major but after one semester transferred to Oberlin College and graduated with a B.A. in mathematics. He then studied at Yale, where he lived in a mathematical commune that listed itself in the New Haven phonebook as N. Bourbaki, and received a Ph.D. in several complex variables. After teaching at the University of Texas at Austin, Kenyon College, and Oberlin College, he is now Professor and Chair of the Department of Mathematical Sciences at Ohio Wesleyan University. He still enjoys both music and mathematics.

ISRAEL KLEINER is professor in the Department of Mathematics and Statistics at York University. He received his Ph.D. at McGill University in ring theory. His interests are in the history of mathematics, in mathematics education (in a broad sense), and in their interface.

Equations for Excellence

The Pleasures of Counting

T.W. Korner



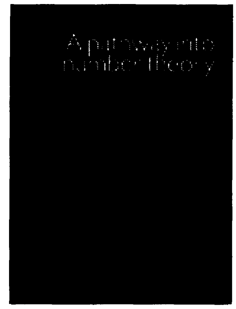
Ranging from the design of anchors and the Battle of the Atlantic to the outbreak of cholera in Victorian Soho, this text describes a variety of lively topics that continue to intrigue professional mathematicians.

1996	544 pp.	
56087-X	Hardback	$85.00
56823-4	Paperback	$34.95

A Pathway Into Number Theory

Second Edition

R.P. Burn



Now in its second edition, this book consists of a sequence of exercises that will lead readers from simple number work to the point where they can prove algebraically the classical results of elementary number theory for themselves.

1996	267 pp.	
57540-0	Paperback	$27.95

A Primer of Probability Logic

Ernest W. Adams

This is a clear well-written text on the subject of probability logic, suitable for advanced undergraduates or graduates, but also of interest to professional philosophers.

CSLI Lecture Notes

1997	c.376 pp.	
1-57586-066-X	Paperback	$24.95

Young Tableaux

With Applications to Representation Theory and Geometry

William Fulton

The aim of this book is to develop the combinatorics of Young tableaux and to show them in action in the algebra of symmetric functions, representations of the symmetric and general linear groups, and the geometry of flag varieties.

London Mathematical Society Student Texts 35

1996	269 pp.	
56144-2	Hardback	$59.95
56724-6	Paperback	$19.95

Mathematical Cavalcade

Brian Bolt

131 activities, ranging from matchstick and coin puzzles through ferrying, railway shunting, dissection, topological and domino problems to a variety of magical number arrays with surprising properties are included.

1992	130 pp.	
42617-0	Paperback	$18.95

Complex Variables

Introduction and Applications

Mark J. Ablowitz
and ***Athanassios S. Fokas***

Part one provides an introduction to the subject, including analytic functions, integration, series, and residue calculus and also includes transform methods, ODEs in the complex plane, and numerical methods. Part two contains conformal mappings, asymptotic expansions, and the study of Riemann-Hilbert problems.

Cambridge Texts in Applied Mathematics 16

1996	c.450 pp.	
48523-1	Paperback	$34.95

The Principles of Mathematics Revisited

Jaakko Hintikka

This book, written by one of philosophy's preeminent logicians, argues that many of the basic assumptions common to logic, philosophy of mathematics and metaphysics are in need of change. Hintikka proposes a new basic first-order logic and uses it to explore the foundations of mathematics.

1996	300 pp.	
49692-6	Hardback	$59.95

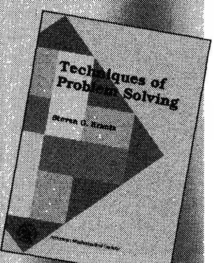
Available in bookstores or from

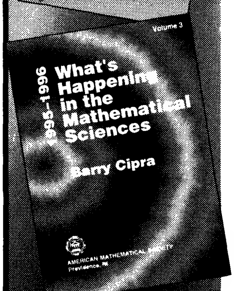
CAMBRIDGE UNIVERSITY PRESS

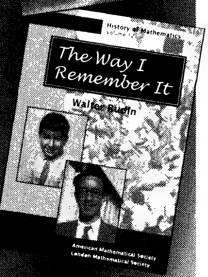
40 West 20th Street, New York, NY 10011-4211
Call toll-free 800-872-7423 MasterCard/VISA accepted.
Prices subject to change. Web site: http://www.cup.org

American Mathematical Society

Mathematics Awareness Week April 20–26, 1997







Groups and Symmetry: A Guide to Discovering Mathematics

David W. Farmer, *Bucknell University, Lewisburg, PA*

Knots and Surfaces: A Guide to Discovering Mathematics

David W. Farmer, *Bucknell University, Lewisburg, PA,* and **Theodore B. Stanford,** *University of Nevada, Reno*

The book is perfectly suited to a course for non-science majors in need of fulfilling a math requirement. All the sections have worked well at sparking student interest and convincing them that math is much more interesting than mere number-crunching and graphing.

—***William Bloch, Wheaton College***

In most mathematics textbooks, the most exciting part of mathematics —the process of invention and discovery—is completely hidden from the reader. The aim of *Groups and Symmetry* and *Knots and Surfaces* is to change all that. By means of a series of carefully selected tasks, these books lead readers to discover some real mathematics. There are no formulas to memorize; no procedures to follow. The books are guides: their job is to start you in the right direction and to bring you back if you stray too far. Discovery is left to you.

Mathematical World, Volume 5 (Farmer); 1996; 102 pages; ISBN 0-8218-0450-2; List $19; All AMS members $15; Order code MAWRLD/5MAA97

Mathematical World, Volume 6 (Stanford); 1996; 101 pages; Softcover; ISBN 0-8218-0451-0; List $19; All AMS members $15; Order code MAWRLD/6MAA97

How to Teach Mathematics: a personal perspective

Steven G. Krantz, *Mathematical Sciences Research Institute, Berkeley, CA*

... an original contribution to the educational literature on teaching mathematics at the post-secondary level. The book itself is an explicit proof of the author's claim "teaching can be rewarding, useful, and fun."

—***Zentralblatt für Mathematik***

1993; 76 pages; Softcover; ISBN 0-8218-0197-X; List $15; All AMS members $12; Order code HTMMAA97

Mathematics and Sports

L. E. Sadovskiĭ and **A. L. Sadovskiĭ**

... a nice survey of applications of mathematics in sporting events.

—***Mathematical Reviews***

Treatment is concise and insightful.

—***Zentralblatt für Mathematik***

This unique book presents simple mathematical models of various aspects of sports, with applications to sports training and competitions. Requiring only a background in precalculus, it would be suitable as a textbook for courses in mathematical modeling and operations research at the high school or college level. Coaches and those who participate in sports will find it interesting as well. The lively writing style and wide range of topics make this book especially appealing.

Mathematical World, Volume 3; 1994; 152 pages; Softcover; ISBN 0-8218-9500-1; List $19; All AMS members $16; Order code MAWRLD/3MAA97

Techniques of Problem Solving

Steven G. Krantz, *Washington University, St. Louis, MO*

The purpose of this book is to teach the basic principles of problem solving, including both mathematical and nonmathematical problems. This book will help students to ...

- translate verbal discussions into analytical data.
- learn problem-solving methods for attacking collections of analytical questions or data.
- build a personal arsenal of internalized problem-solving techniques and solutions.
- become "armed problem solvers", ready to do battle with a variety of puzzles in different areas of life.

1996; 465 pages; Softcover; ISBN 0-8218-0619-X; List $29; All AMS members $23; Order code TPSMAA97

The Way I Remember It

Walter Rudin, *University of Wisconsin, Madison*

Walter Rudin's memoirs should prove to be a delightful read specifically to mathematicians, but also to historians who are interested in learning about his colorful history and ancestry. As those who are familiar with Rudin's writing will recognize, he brings to this book the same care, depth, and originality that is the hallmark of his work.

Co-published with the London Mathematical Society. Members of the LMS may order directly from the AMS at the AMS member price. The LMS is registered with the Charity Commissioners.

History of Mathematics, Volume 12; 1996; 191 pages; Hardcover; ISBN 0-8218-0633-5; List $29; All AMS members $23; Order code HMATH/12MAA97

What's Happening in the Mathematical Sciences, 1995–1996

Barry Cipra

... stylish format ... largely accessible to laymen ... This publication is one of the snappier examples of a growing genre from scientific societies seeking to increase public understanding of their work and its societal value.

—***Science & Government Report***

What's Happening in the Mathematical Sciences, Volume 3; 1996; 111 pages; Softcover; ISBN 0-8218-0355-7; List $12; Order code HAPPENING/3MAA97

All prices subject to change. Charges for delivery are $3.00 per order. For air delivery outside of the continental U. S., please include $6.50 per item. *Prepayment required.* Order from: **American Mathematical Society,** P. O. Box 5904, Boston, MA 02206-5904. For credit card orders, fax (401) 331-3842 or call toll free 800-321-4AMS (4267) in the U. S. and Canada, (401) 455-4000 worldwide. Or place your order through the AMS bookstore at http://www.ams.org/bookstore/. Residents of Canada, please include 7% GST.

SPRINGER FOR MATHEMATICS

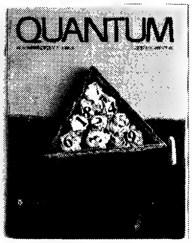
ALEXANDER J. HAHN, University of Notre Dame

Learning Basic Calculus

From Archimedes to Newton to its Role in Science

Part I: FROM ARCHIMEDES TO NEWTON develops calculus, as well as the necessary trigonometry and analytic geometry, from within the relevant historical context, be it that of the Greek thinkers, Galileo, Kepler, Descartes, Leibniz, or Newton.
Part II: CALCULUS AND THE SCIENCES develops the calculus again, this time in a more rigorous way. Comparisons with the approaches of Leibniz and Newton point to the necessity of certain theoretical concerns. But the primary purpose of this part is the illustration of the fact that calculus informs, enlightens, and gives essential substance to, a wide horizon of disciplines of science, engineering, and business.
For much more information visit:http://www.nd.edu:80/~hahn/

1997/APP. 300 PP., 80 ILLUS./HARDCOVER/$44.50 (TENT.)/ISBN 0-387-94606-3
TEXTBOOKS IN MATHEMATICAL SCIENCES



QUANTUM

The Magazine of Math and Science

Since every issue of *Quantum* provides you with the most informative and fun information in math and science, your subscription to *Quantum* is the best way to stimulate, improve, and coordinate your teaching and your students' learning of math and science.
Quantum's regular departments include: *Brainteasers, How Do You Figure?, At the Blackboard, Kaleidoscope,* and *In the Lab,* which provide quick and fun problems and stories to help teach students about different areas of physics and mathematics. *Crisscross Science* presents science crossword puzzles.

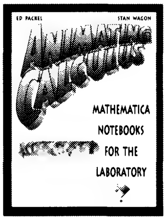
Visit *Quantum*'s Website:
http://www.nsta.org/quantum/index.html/

ISSN 1048-8820 TITLE NO. 583
1997, ONE YEAR SUBSCRIPTION,
6 ISSUES
SUBSCRIPTION RATES:
INSTITUTIONAL RATE: $45.00
SPECIAL PERSONAL RATE: $25.00
SUPER STUDENT RATE: $18.00
(*SOCIETY RATE AVAILABLE,
PLEASE INQUIRE.)

E. PACKEL, Lake Forest College, Lake Forest, IL and
S. WAGON, Macalester College, St. Paul, MN

Animating Calculus

Mathematica Notebooks for the Laboratory



Animating Calculus is the result of an inspired collaboration between Ed Packel, who is experienced in the integration of computers and mathematics in the classroom, and Stan Wagon, a well-known mathematical expositor and author of the acclaimed *Mathematica in Action* (1991, W.H. Freeman). This book contains 22 laboratory notebooks that use Mathematica. With its superior animations and graphics, and coverage ranging from standard topics to new and unusual extensions and applications, *Animating Calculus* is a remarkable tool for experiencing and learning about calculus on the computer.

1996/292 PP./INCLUDES 3.5" DISK/SOFTCOVER/$34.95
ISBN 0-387-94748-5

DAN SCHWALBE and **STAN WAGON,** both of Macalester College, St. Paul, MN

Visual DSolve

Visualizing Differential Equations with Mathematica

This software and manual is a comprehensive Mathematica package for the visualization of solutions to ordinary differential equations. It includes well-known DE visualization tools such as solution plots and orbits, but also includes some new ideas such as the side of shaded gray regions in the phase plane. The package can be used in the traditional differential equations course to enhance students' understanding of solutions of ordinary differential equations.

1997/APP. 288 PP., 316 ILLUS./INCLUDES 3.5" DISK/SOFTCOVER/$34.95
ISBN 0-387-94721-3

Springer
www.springer-ny.com

Order Today!
CALL: 1 800 SPRINGER or
FAX: (201) 348-4505
WRITE: Springer-Verlag New York, Inc., Dept. #S270, PO Box 2485, Secaucus, NJ 07096-2485
VISIT: Your local technical bookstore
E-MAIL: orders@springer-ny.com
INSTRUCTORS: Call or write for info on textbook exam copies.
2/97 REFERENCE S270

THE AMERICAN MATHEMATICAL MONTHLY

Volume 104, Number 4 April 1997

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and Notes should be sent to the Editor:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send your email address and 3 copies of the complete manuscript (including all figures with captions and lettering), typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper.

Letters to the Editor on any topic are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

See the MONTHLY section of MAA Online for current information such as contents of issues, descriptive summaries of forthcoming articles, tips for authors, and preparation of manuscripts in TEX:

http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all material, typewritten on only one side of the paper.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Mr. JOE WATSON, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Rectangular Invertible Matrices

A. J. Berrick and M. E. Keating

In everyday mathematics, a matrix with a two-sided inverse must be a square matrix. However, there are situations in which rectangular "invertible" matrices do occur, and our purpose in this note to give a brief survey of such situations and the matrices that arise in them.

The usual rule for the multiplication of real matrices works equally for matrices with entries in any ring R, and we say that an $m \times n$ matrix P, having entries in R, is *invertible* if there is a matrix Q with entries in R so that

$$PQ = I_m \qquad \text{and} \qquad QP = I_n,$$

where I_m and I_n are identity matrices of the correct sizes. The matrix Q is called the *inverse* of P. If it exists, it must be an $n \times m$ matrix and it will be unique.

Whether or not an invertible matrix is necessarily square is a property that depends on the ring from which the entries of our matrices are selected. If every invertible matrix with entries in R must be square, then R is said to have *invariant basis number*, a term that we justify shortly. Familiar rings, such as fields or the integers, do enjoy this property, and some fundamental results show that the property can often be transmitted from one ring to another. A consequence is that rings with invariant basis number form a large class of rings, encompassing most of those encountered by a working mathematician.

To redress the balance, we give some examples of rings that do not have invariant basis number, and we show that they can be classified in terms of a congruence relation on the set of natural numbers, which tells us the permitted sizes of invertible matrices for that ring. Such a congruence relation is in turn determined by its "type" (w, d), which may then be called the type of the ring.

This classification of rings by type is useful in determining whether there can be a ring homomorphism between rings of differing types, and leads to the question of the possibility of representing a ring of one type as a subring of a ring of another type.

Our results (apart from applications) are not new, and our derivation of the theory of congruences on $\mathbb{N}$ is elementary, but there does not seem to be any account that contains this result in conjunction with a discussion of rings without invariant basis number.

All rings in this paper are associative and possess an identity element; ring homomorphisms must preserve identity elements. We take the natural numbers to be $\mathbb{N} = \{1, 2, \dots\}$.

1. We start with a justification of the terminology and a survey of results concerning invariant basis number and the lack of it.

For any integer $n \geq 1$, we write R^n for the space of column vectors of length n with entries in R; more formally, R^n is the *free right R-module of rank n*. As in the theory of vector spaces, the standard unit vectors $e_1, \dots, e_n$ are linearly independent and span R^n, that is, they form a *basis* of R^n.

Now let P be an $m \times n$ matrix with entries in R. There is a linear transformation from R^n to R^m defined by the rule $x \mapsto Px$, and, again as in vector space theory, if P is invertible, the vectors $Pe_1, \ldots, Pe_n$ then form a basis of R^m. However, R^m already has a basis with m members, so if the number of elements in a basis of R^m is invariant, we must have $m = n$, that is, our invertible matrix has to be square.

The argument also works in the reverse direction: if we do have a basis of R^m with n members, we can find an invertible $m \times n$ matrix P that gives the members of the basis in the form $Pe_1, \ldots, Pe_n$.

Thus we see why the expression "R has invariant basis number" is synonymous with the assertion that an invertible matrix over R must be square. Henceforth we abbreviate invariant basis number to *IBN*.

Our discussion also shows that a field must have IBN, since the number of elements in a basis of a vector space is just its dimension, which we know to be unique.

Further examples of rings that have IBN are provided by the observation that if $f\colon R \to S$ is a ring homomorphism and P is an invertible matrix with entries in R, having inverse Q, then the image matrix fP is an invertible matrix with entries in S, with inverse fQ. Thus, if S has IBN so also has R.

For example, suppose that R is commutative. Then it is a fact that there is a homomorphism $R \to F$ for some field F ([**1**], p. 3) and so R has IBN. This conclusion can also be proved by a direct calculation with matrices ([**6**], §10.4).

A deeper result is that (right) Noetherian rings have IBN ([**4**], 4.3, Theorem 7). Thus, "naturally occurring" rings tend to have invariant basis number.

The easiest example of a ring without IBN is provided by the *cone* $C(R)$ of a ring R. The elements of $C(R)$ are the infinite matrices over R, with rows and columns indexed by the natural numbers, each row and column having only a finite number of nonzero entries. The elements

$$\beta = \begin{pmatrix} 1\,0\,0\,0\,0\,0\cdots \\ 0\,0\,1\,0\,0\,0\cdots \\ 0\,0\,0\,0\,1\,0\cdots \\ \cdots \end{pmatrix} \quad \text{and} \quad \gamma = \begin{pmatrix} 0\,1\,0\,0\,0\,0\cdots \\ 0\,0\,0\,1\,0\,0\cdots \\ 0\,0\,0\,0\,0\,1\cdots \\ \cdots \end{pmatrix}$$

of $C(R)$ give an invertible 2×1 matrix $\alpha = \begin{pmatrix} \beta \\ \gamma \end{pmatrix}$, with inverse the transpose matrix $(\beta^t\,\gamma^t)$.

Further examples of rings without IBN are not so easy to find. Some can be constructed as follows. Choose (unequal) natural numbers m and n, and let $X = (X_{hi})$, $m \times n$, and $Y = (Y_{jk})$, $n \times m$, be matrices whose entries are non-commuting indeterminates. We can then form a noncommutative polynomial ring (otherwise known as a free associative algebra) $K\langle X, Y\rangle$ by adjoining the $2mn$ entries of the matrices X and Y to a field K. Now let the *Leavitt ring* $L(m,n)$ be the quotient ring of $K\langle X, Y\rangle$ obtained by equating to 0 all the entries of the matrices $XY - I_m$ and $YX - I_n$. By design, the ring $L(m,n)$ has a pair of mutually inverse rectangular matrices, of sizes $m \times n$ and $n \times m$, namely the images of X and Y respectively. What is not at all clear is that $L(m,n)$ is non-trivial and that it has no smaller pair of mutually inverse rectangular matrices. Proofs of these assertions are given in [**10**] and [**11**], and, using different methods, in [**3**] and [**2**].

2. Given a ring without IBN, it is natural to ask if there is a non-square invertible matrix over it of minimal size, and, if so, whether the possible sizes of non-square

invertible matrices can be determined in terms of this minimal size. The answer to both questions is in the affirmative, which fact leads to a classification of rings without IBN.

The questions are tackled by associating with any ring R, having IBN or not, a relation $\sim_R$ on the set $\mathbb{N}$ of natural numbers that is defined by the existence of invertible matrices.

Given $a, b \in \mathbb{N}$, we say that $a \sim_R b$ if and only if there is an $a \times b$ invertible matrix P over R. The relation $\sim_R$ is clearly reflexive, symmetric (consider the matrix inverse to P), and transitive (if the $b \times c$ matrix Q is invertible, so is PQ). Thus $\sim_R$ is an equivalence relation.

The relation $\sim_R$ has the further property that it is also a *congruence* on $\mathbb{N}$. By definition, such a relation $\sim$ is an equivalence relation in the ordinary sense, and it satisfies the extra, additivity, property that if $a \sim b$ and $c \sim d$, then $a + c \sim b + d$ also.

To see that $\sim_R$ is a congruence, note that if P is an $a \times b$ invertible matrix and Q is a $c \times d$ invertible matrix, then the block matrix $\begin{pmatrix} P & O \\ O & Q \end{pmatrix}$ is also invertible, of size $(a + c) \times (b + d)$.

When R has IBN, the relation $\sim_R$ is simply equality.

3. Now we turn to the description of a general congruence on $\mathbb{N}$. We start by relating congruences on $\mathbb{N}$ with the more familiar congruences on the set $\mathbb{Z}$ of all integers. If we replace $\mathbb{N}$ by $\mathbb{Z}$ in the definition, it is easy to see that a congruence as defined here is simply congruence mod d in the usual sense, where d is the smallest natural number congruent to 0 if the congruence is not equality; d is called the *modulus* of the congruence.

Suppose that $\sim$ is a congruence on $\mathbb{N}$. Then we can extend $\sim$ to a congruence $\equiv$ on $\mathbb{Z}$ by saying that $a \equiv b$ if there is a natural number x so that

$$a + x \sim b + x.$$

It is routine to verify that $\equiv$ is in fact a congruence on $\mathbb{Z}$. So, if $\sim$ is not equality, we can associate with $\sim$ the modulus d of the extension congruence $\equiv$ on $\mathbb{Z}$.

To avoid confusion, we shall always use $\sim$ for a congruence on $\mathbb{N}$ and $\equiv$ for a congruence on $\mathbb{Z}$.

Working in the opposite direction, given any natural number d we can obtain a congruence $\sim$ on $\mathbb{N}$ by restricting to $\mathbb{N}$ an ordinary congruence mod d on the set $\mathbb{Z}$. For such a congruence on $\mathbb{N}$, the congruence classes are

$$\{e, e + d, e + 2d, \dots\}, \quad e = 1, \dots, d.$$

In general, a congruence relation on $\mathbb{N}$ is a mixture of equality and a restriction of a congruence mod d from $\mathbb{Z}$. More precisely, we prove the following result.

Theorem 1. *Let $\sim$ be a congruence relation on $\mathbb{N}$ that is not equality. Then there are unique natural numbers w and d so that the set of equivalence classes under $\sim$ comprises $w - 1$ classes*

$$\{1\}, \dots, \{w - 1\},$$

each of which has a single element, together with d infinite classes

$$\{e, e + d, e + 2d, \dots\}, \quad e = w, \dots, w + d - 1.$$

Conversely, any such partition of $\mathbb{N}$ gives a congruence relation on $\mathbb{N}$.

The number w is the least natural number congruent to a different natural number under $\sim$, and the number d is the modulus of the extension of $\sim$ to $\mathbb{Z}$.

The pair (w, d) is called the type *of the congruence.*

Proof: It is straightforward to check that a partition of $\mathbb{N}$ as in the assertion does define a congruence on $\mathbb{N}$.

Working in the other direction, suppose $\sim$ to be a congruence relation on $\mathbb{N}$ that is not equality, and let d be the modulus of the extension congruence on $\mathbb{Z}$. Then there exist natural numbers x with $x \sim x + d$. So define w to be the least such x. Clearly, for any $e \geq w$,

$$e = w + (e - w) \sim w + d + (e - w) = e + d,$$

so the equivalence class of e contains

$$\{e, e + d, e + 2d, \dots\}.$$

For $u < w$, we show that the equivalence class of u is a singleton. Otherwise, by the additivity property it is infinite and we may take v to be its least member exceeding $w - 1$. So by the previous paragraph, $v \sim v + d$. Then from $u \sim v$ we have $u + d \sim v + d$. Thus by transitivity $u \sim u + d$, contradicting the definition of w.

Thus the singleton classes are

$$\{1\}, \dots, \{w - 1\}$$

and the remaining classes are as claimed. ∎

Remarks

(a) The congruences on $\mathbb{N}$ that are obtained by restriction from a congruence on $\mathbb{Z}$ are those of type $(1, d)$.

(b) It is convenient to give equality the type $(\infty, 0)$.

(c) The theorem is known to semigroup theorists, but there does not seem to be any account of it in elementary texts on the number system. It can be deduced from general results on semigroups given in several texts, for example, [**8**], §I.1.2 and Theorem 5.3. The theorem is stated without proof by Cohn in [**3**]; he also gives an elementary proof in Lemma X.3.1 of [**5**] in the midst of a more sophisticated discussion.

4. A ring R said to have the type (w, d) of the congruence $\sim_R$ on $\mathbb{N}$ defined in Section 2. This gives us our desired classification of rings that do not have IBN. Thus the rings $L(m, n)$ have type $(m, n - m)$ (taking $m < n$) and the cone of any ring has type $(1, 1)$.

Notice that the rings with IBN are precisely those of type $(\infty, 0)$, since this is the type of the equality congruence.

As an application of the notion of the type of a ring, we compare the types of rings R and S when there is a ring homomorphism $f\colon R \to S$. If $a \sim_R b$, then $a \sim_S b$ also, since the images of mutually inverse matrices are again mutually inverse. Thus the relation $\sim_R$ is *finer* than $\sim_S$.

It follows that each $\sim_S$-equivalence class is a disjoint union of $\sim_R$-equivalence classes. Therefore, if (w, d) and (u, c) are the respective types of R and S, then $u \leq w$ and c divides d. (Here, we take the symbol ∞ to exceed all natural numbers.)

In the reverse direction, if (w, d) and (u, c) are the respective types of abstract congruences $\sim$ and $\approx$ on $\mathbb{N}$, and if $u \leq w$ and c divides d, then it is easy to see that $\sim$ is finer than $\approx$.

5. As an application, we recover a result that we proved in Section 1: if there is a homomorphism $f\colon R \to S$ and S has invariant basis number, so also has R. Take (w, d) to be the type of R. Since S has type $(\infty, 0)$, we have $\infty \leq w$ and $0|d$, which forces the equalities $\infty = w$ and $0 = d$.

In particular, if S is either commutative or Noetherian, then R has IBN.

We can also give a reason why equality and rings with IBN must be assigned the type $(\infty, 0)$. On the one hand, a field K has IBN, and, on the other hand, for any pair of natural numbers u and c, K can be embedded in the Leavitt ring $L(u, u+c)$ of type (u, c). Thus the type (w, d) of K must have the properties that w exceeds every natural number and that d is divisible by every natural number.

6. Here are some exercises for readers who know some ring theory.

(a) Suppose that $D = R_1 \times R_2$, the direct product of rings, and that R_1 and R_2 have types (w_1, d_1) and (w_2, d_2) respectively. Then D has type $(\max(w_1, w_2), \mathrm{lcm}(\mathrm{d}_1, \mathrm{d}_2))$.

(b) If R has type $(w, d) \neq (\infty, 0)$, then the ring $M_n(R)$ of $n \times n$ matrices over R has type $(\langle w/n \rangle, d/\gcd(d, n))$, where $\langle w/n \rangle$ is the least integer with $\langle w/n \rangle \geq w/n$. Thus if n is a large multiple of d, $M_n(R)$ has type $(1, 1)$.

Note that the type of a ring is not preserved under Morita equivalence of rings (see **[4]**, §4.5, for a definition of this term).

(c) Suppose that a ring R is the union $R = \bigcup_i R_i$ of a set of subrings. Then there is a natural number N so that the rings R_i for $i \geq N$ all have the same type, which is the type of R. The corresponding statement also holds for more general direct limits of rings.

Inverse limits seem to be less well behaved.

(d) Let $\mathrm{rad}(R)$ be the Jacobson radical of R, that is, the intersection of all the maximal right ideals of R. Then R has the same type as $R/\mathrm{rad}(R)$.

Hint: It suffices to show that if an $n \times k$ matrix X over $R/\mathrm{rad}(R)$ is invertible, then there is some invertible $n \times k$ matrix over R that maps to X under the canonical surjection from R to $R/\mathrm{rad}(R)$. In fact, any $n \times k$ matrix A over R that maps to X must be invertible. To see this, regard A as an R-module homomorphism from R^k to R^n of free right R-modules. Since

$$X(R/\mathrm{rad}(R))^k = (R/\mathrm{rad}(R))^n,$$

we have

$$AR^k + (\mathrm{rad}(R))^n = R^n.$$

By Nakayama's Lemma (**[4]**, §10.3), $AR^k = R^n$, that is, A is a surjection. Thus there is a $k \times n$ matrix B with $AB = I_n$. But B must map to the unique inverse of X, so the same argument shows that B has an $n \times k$ right inverse, which must be A again.

7. There is no reason why there should be a homomorphism from a ring R to a ring S even when $\sim_R$ is finer than $\sim_S$. For example, there are no homomorphisms between the finite fields $\mathbb{F}_p$ and $\mathbb{F}_q$ if p and q are different primes, although they have the same type.

An interesting question is the extent to which rings of one type can usefully be represented in terms of rings of another type. For instance, the cone $C(K)$ over a field K, which has "terminal" type (1, 1), contains subrings of all other types – this follows from Proposition 2.1 of **[7]**, since the Leavitt rings over K all have countable dimension. However, it is not clear how the properties of a Leavitt ring can be recovered from its embedding in the cone.

A result in the reverse direction is given in **[12]**.

8. A ring R is said to have *stable rank* 1 (sr $R = 1$) if for any pair of elements a_0, a_1 such that $R = Ra_0 + Ra_1$, there exists $b \in R$ with $ba_0 + a_1$ a unit in R. In such a ring, the relation $xy = 1$ always yields $yx = 1$ (after $a_0 = 1 - yx$, $a_1 = x$) **[9]**. Moreover, if sr $R = 1$, then for all n we have sr $M_n(R) = 1$ **[13]**. So suppose that P is an $m \times n$ matrix over R with sr $R = 1$, and that $PQ = I_m$ for some matrix Q over R. By adjoining $n - m$ zero columns to P, we obtain $\overline{P} \in M_n(R)$, which has a right inverse $\overline{Q}$. Since sr $M_n(R) = 1$, $\overline{Q}$ must also be left inverse to $\overline{P}$, in contradiction of the fact that $\overline{QP}$ must contain at least $n - m$ zero columns. It follows that a ring R with stable rank 1 has IBN.

ACKNOWLEDGMENTS. We thank Professor J. M. Howie for his advice on semigroup theory, and the referees for helpful comments on earlier versions of this note.

REFERENCES

1. M. F. Atiyah and I. G. MacDonald, *Introduction to Commutative Algebra*, Addison-Wesley, Reading, Mass. 1969.
2. G. M. Bergman, Coproducts and some universal ring constructions, *Trans. Amer. Math. Soc.* **200** (1977), 33–88.
3. P. M. Cohn, Some remarks on the invariant basis property, *Topology* **5** (1966), 215–228.
4. ______, *Algebra II*, Wiley & Sons, Chichester, 1977.
5. ______, *Universal Algebra*, Reidel, Dordrecht, 1981.
6. ______, *Algebra I*, Wiley & Sons, Chichester, 1982.
7. K. Goodearl, P. Menal, and J. Moncasi, Free and residually artinian regular rings, *J. Algebra* **156** (1993), 407–432.
8. J. M. Howie, *An Introduction to Semigroup Theory*, London Math. Soc. Monograph 7, Academic Press, London, 1976.
9. W. van der Kallen, Injective stability for K_2, *Lecture Notes in Math.* **551**, Springer, Berlin, 1976, pp. 77–154.
10. W. G. Leavitt, Modules without invariant basis number, *Proc. Amer. Math. Soc.* **8** (1957), 322–328.
11. ______, The module type of a ring, *Trans. Amer. Math. Soc.* **103** (1962), 113–130.
12. ______, The module type of a homomorphic image, *Duke Math. J.* **32** (1965), 305–311.
13. L. N. Vaserstein, Stable ranks of rings and dimensionality of topological spaces, *Funct. Anal. and Appl.* **5** (1971), 102–110.

Department of Mathematics
National University of Singapore
Singapore 119260
berrick@math.nus.sg

Department of Mathematics
Imperial College
London SW7 2BZ
England UK
m.keating@ic.ac.uk

Equal Pay for All Prisoners

Maarten C. Boerlijst, Martin A. Nowak, and Karl Sigmund

By prisoners we mean, of course, players of the well-known Prisoner's Dilemma game (to be described presently). We shall show that there exist simple strategies for the infinitely iterated Prisoner's Dilemma that act as *equalizers* in the sense that all co-players receive the same payoff, no matter what their strategies are like.

The Prisoner's Dilemma game, a favorite with game theorists, social scientists, philosophers, and evolutionary biologists, displays the vulnerability of cooperation in a minimalistic model (see [1] to [5]). The two players engaged in this game can choose whether to cooperate or to defect. If both defect, they gain 1 point each; if both cooperate, they gain 3 points; but if one player defects and the other does not, then the defector receives 5 points and the other player only 0. The right move is obviously to defect, no matter what the other player does. As a result, both players earn 1 point instead of 3.

But if the same two players repeat the game very frequently, there exists no strategy that is best against all comers. The diversity of strategies is staggering. If we simulate on a computer populations of strategies evolving under a mutation-selection regime (with mutation introducing new strategies and selection weening out those with lowest payoff), we observe a rich variety of evolutionary histories frequently leading to cooperative regimes dominated by strategies like Pavlov (cooperate whenever the opponent's move, in the previous round, matched yours) or Generous Tit For Tat (always reciprocate your opponent's cooperative move, but reciprocate only two-thirds of the defections). Remarkably, all strategies of the iterated Prisoner's Dilemma, which can be very complex and make up a huge set, obtain the same payoff against some rather simple equalizer strategies.

More generally, let us consider a two-player game where both players have the same two strategies and the same payoff matrix. We denote the first strategy (row 1) by **C** (for 'cooperate') and the second (row 2) by **D** (for 'defect') and write the payoff matrix as

$$\begin{array}{ccc|cc} & & & \multicolumn{2}{c}{\text{Opponent}} \\ & & & \mathbf{C} & \mathbf{D} \\ \hline \text{You} & & \mathbf{C} & R,R & S,T \\ & & \mathbf{D} & T,S & P,P \end{array} . \qquad (1)$$

Such games include the Prisoner's Dilemma, where $T > R > P > S$, and the Chicken game, where $T > R > S > P$. (In the Prisoner's Dilemma case, R stands for the *reward* for mutual cooperation, P is the *penalty* for mutual defection, T is the *temptation* payoff for unilaterally defecting and S the *sucker* payoff for being exploited.)

Let us assume that the game is repeated infinitely often. A *strategy* in such a supergame is a program telling the player in each round whether to play **C** or **D**. The program may be history-dependent and stochastic: it specifies at every step the probability for playing **C**, depending on what happened so far. If A_n is the

payoff in the n-th round, the expected long-run average payoff for a player is given by

$$\lim_{N\to\infty} \frac{A_1 + \cdots + A_N}{N}, \tag{2}$$

provided it exists. It need not always exist: think of two players cooperating in the first 10 rounds, defecting in the next 100 rounds, then cooperating in the following 1000 rounds, etc.

Memory-one strategies are particularly simple. Such a strategy is given by the probability to play **C** in the first round, and a quadruple $\mathbf{p} = (p_R, p_S, p_T, p_P)$, where p_i denotes the probability that the player plays **C** after having experienced outcome $i \in \{R, S, T, P\}$ in the previous round. Some of the most successful strategies belong to this class, including Generous Tit For Tat $(1, 1/3, 1, 1/3)$ and Pavlov $(1, 0, 0, 1)$.

Theorem. *If* $\max(S, P) < \min(R, T)$, *then there exist, for every value* π *between these numbers, memory-one strategies* $\mathbf{p}$ *such that every opponent obtains the long-run average payoff* π *against a player using such a strategy. The vector* $\mathbf{p}$ *is given by*

$$(1 - (R - \pi)a, 1 - (T - \pi)a, (\pi - S)a, (\pi - P)a) \tag{3}$$

where a *is any real number such that* $1/a \geq \max(T - \pi, R - \pi, \pi - S, \pi - P)$.

Proof: The condition on a guarantees that the p_i are probabilities. Let us denote by $q_i(n)$ the conditional probability that the opponent plays **C** in the following round, given that the n-th round resulted in outcome i, and by $s_i(n)$ the probability that the outcome in the n-th round is i. By conditioning on round n, we obtain:

$$\begin{aligned} s_R(n+1) = s_R(n)q_R(n)[1 - (R - \pi)a] + s_S(n)q_S(n)[1 - (T - \pi)a] \\ + s_T(n)q_T(n)(\pi - S)a + s_P(n)q_P(n)(\pi - P)a. \end{aligned} \tag{4}$$

Similarly,

$$\begin{aligned} s_S(n+1) &= s_R(n)(1 - q_R(n))[1 - (R - \pi)a] \\ &+ s_S(n)(1 - q_S(n))[1 - (T - \pi)a] \\ &+ s_T(n)(1 - q_T(n))(\pi - S)a + s_P(n)(1 - q_P(n))(\pi - P)a. \end{aligned} \tag{5}$$

Summing (4) and (5) yields the probability that you play **C** in round $n + 1$

$$\begin{aligned} s_R(n+1) + s_S(n+1) = s_R(n)[1 - (R - \pi)a] + s_S(n)[1 - (T - \pi)a] \\ + s_T(n)(\pi - S)a + s_P(n)(\pi - P)a. \end{aligned}$$

Hence

$$\begin{aligned} &a^{-1}[s_R(n) + s_S(n) - s_R(n+1) - s_S(n+1)] = \\ &Rs_R(n) + Ss_T(n) + Ts_S(n) + Ps_P(n) - \pi[s_R(n) + s_S(n) + s_T(n) + s_P(n)]. \end{aligned} \tag{6}$$

Since the $s_i(n)$ sum up to 1, the right-hand side is just $A_n - \pi$, where A_n is the *opponent's* payoff in the n-th round (we must bear in mind that one player's outcome S is the other player's outcome T). Summing up (6) for $n = 1, \ldots, N$ and dividing by N, we obtain

$$\frac{1}{aN}[s_R(1) + s_S(1) - s_R(N+1) - s_S(N+1)] = \frac{A_1 + \cdots + A_N}{N} - \pi,$$

and hence

$$\lim_{N\to\infty} \frac{A_1 + \cdots + A_N}{N} = \pi. \qquad \blacksquare$$

A few final remarks. Two players using equalizer strategies are in *Nash equilibrium*, which means that neither has an incentive to change strategy. Nash equilibria exist for every game; for iterated games, they abound. Indeed, the so-called Folk Theorem in game theory states that every feasible pair of payoff-values exceeding the *minimax* (the highest payoff that a player can enforce, which in our case is $\max(S, P)$) can be realized by a Nash-equilibrium pair [**2**, p. 373]. Our theorem is related to this: the strategies are equalizers with memory one. Two players using such strategies have no reason to switch unilaterally to another strategy, since they cannot improve their payoff; however, they have no reason *not* to adopt another strategy either, since they will not be penalised. Since their opponent plays an equalizer strategy, they can switch to *any* other strategy, and not be worse off. If *both* players opt for a change, however, they are likely to end up in a non-equilibrium situation.

If a is chosen small enough, the runs of consecutive defections or cooperations can be made arbitrarily long. The condition $\min(R, T) > \max(S, P)$ and its converse are not only sufficient, but also necessary for the existence of such equalizer strategies. It is easy to construct other equalizer strategies. For example, play **C** until the opponent's mean payoff is larger than π, then play **D** until it is smaller than π, then play **C** until it is larger again, etc. However, such a strategy requires monitoring the opponent's entire payoff sequence. The point is that even within memory-one strategies, equalizers exist.

ACKNOWLEDGMENTS. Financial support from the Wellcome Trust (MAN) and the Austrian Forschungsförderungsfonds (KS) is gratefully acknowledged.

REFERENCES

1. R. Axelrod, *The Evolution of Cooperation*, Penguin, Harmondsworth, 1990.
2. K. G. Binmore, *Fun and Games: a Text on Game Theory*, Heath and Co, Lexington, Massachusetts, 1992.
3. M. Mesterton-Gibbons, *An Introduction to Game-theoretic Modelling*, Addison-Wesley, Redwood City, California, 1992.
4. M. A. Nowak, R. M. May, and K. Sigmund, The Arithmetics of Mutual Aid, *Scientific American* **272** (1995), 76–81.
5. K. Sigmund, *Games of Life: Explorations in Ecology, Evolution and Behaviour*, Penguin, Harmondsworth, 1995.

Maarten C. Boerlijst and Martin A. Nowak
Department of Zoology
University of Oxford
South Parks Road, Oxford OX1 3PS
UK
mcb@einstein.zoo.ox.ac.uk
nowak@nowak.zoo.ox.ac.uk

Karl Sigmund
Institut für Mathematik
Universität Wien
Strudlhofgasse 4, A-1090 Vienna
Austria
ksigmund@esi.ac.at

The Long and the Short on Counting Sequences

Jim Sauerberg and Linghsueh Shu

1. INTRODUCTION. Consider the sequence of positive integers $S_0 = 2, 1, 1, 4$. S_0 consists of two 1's, one 2, and one 4, so let us define S_1 to be this description: $S_1 = 2, 1, 1, 2, 1, 4$. Repeating this process, S_1 consists of three 1's, two 2's and one 4, so set $S_2 = 3, 1, 2, 2, 1, 4$. Continuing in this way for several more steps produces

$$S_3 = 2, 1, 2, 2, 1, 3, 1, 4$$

$$S_4 = 3, 1, 3, 2, 1, 3, 1, 4$$

$$S_5 = 3, 1, 1, 2, 3, 3, 1, 4$$

$$S_6 = 3, 1, 1, 2, 3, 3, 1, 4.$$

In general, given any finite sequence of positive numbers S_0, this process of constructing S_{i+1} to be the sequence that counts how many times each number in S_i appears in S_i creates a *counting sequence* $\{S_i\}_{i \geq 0}$. As the reader certainly noticed, in our counting sequence we have $S_5 = S_6 = S_7 = \cdots$. In fact, in any counting sequence, because S_{i+1} is uniquely determined by S_i, if there exist numbers p and i such that $S_i = S_{i+p}$, then $S_{i'} = S_{i'+p}$ for all $i' \geq i$. We then say that $\{S_i\}_{i \geq 0}$ is *ultimately periodic*. The rather surprising main result of **[1]** is

Theorem 1. *For any finite sequence of positive integers S_0, the associated counting sequence $\{S_i\}_{i \geq 0}$ is ultimately periodic. In other words, given S_0 there are integers p_0 and p so that $S_{i+p} = S_i$ for all $i \geq p_0$.*

The smallest p_0 and smallest p satisfying Theorem 1 are called the *pre-period* and the *period* of the counting sequence $\{S_i\}$. Then a *periodic counting sequence* of period p, or simply a *p-cycle*, is a counting sequence of pre-period 0 and period p. For example, the counting sequence corresponding to $S_0 = 2, 1, 1, 4$ has pre-period 5 and period 1, that is, it "ends" in a 1-cycle. Similarly, the counting sequence corresponding to $S_0 = 5, 6$ ends in a two-cycle, and the counting sequence corresponding to $S_0 = 6, 7$ ends in a three-cycle.

Several different types of counting sequences have been studied in recent years (see **[1]**, **[5]**, **[6]**, **[7]**, **[8]**, and M4779 in **[9]**). In this paper we consider these counting sequences, bring out their connections, and explore the periodic behavior of each. To expand on this, the questions we answer are:

1) What are the possible periods p? For each p, how many p-cycles are there? In Section 3 we find all possible periods and classify all cycles. Partial answers to these questions are given in **[6]**.
2) A puzzle of Raphael Robinson **[3**, pp. 389–390**]** asks the reader to place numbers in the blanks so that the following is true: "In this sentence, the number of occurrences of 0 is __, of 1 is __, of 2 is __, of 3 is __, of 4 is __, of 5 is __, of 6 is __, of 7 is __, of 8 is __, and of 9 is __." To find such a

sentence we must find a one-cycle that contains all of the numbers in base 10, as opposed to the infinite base consisting of all the natural numbers implicitly used in the preceding paragraphs. More generally, one can build counting sequences in base k for any $k \geq 2$. Are such counting sequences also eventually periodic? In Section 4 we show that they are, and determine exactly how many different cycles there are in each base. This expands upon the results of **[6]**.

3) What happens when S_0 is replaced by an infinite sequence? It is very easy to give infinite sequences S_0 such that $\{S_i\}_{i \geq 0}$ is not well-defined. In Section 5 we show how to construct examples of infinite sequences S_0 so that $\{S_i\}_{i \geq 0}$ is well-defined and is ultimately periodic. We also give two different methods for constructing infinite sequences S_0 so that $\{S_i\}_{i \geq 0}$ is well-defined and but not ultimately periodic.
4) The second term, fourth term, sixth term, etc., in each sequence S_i of a counting sequence do little more than serve as place holders. Assuming there is a way to tell which integer each number is describing, what happens if we form counting sequences without these place holders? One can then ask questions similar to those in 1) for these sequences. These questions have, for the most part, been answered in **[5]**, **[7]**, and **[8]**. We see in Section 6 that the answers also follow as very simple corollaries of our work in Sections 2 and 3. Robinson's question can also be asked in this context; see **[10]**.

In each of the various methods we use to construct counting sequences, the successor sequence lists the number of appearances of a particular digit throughout the entire previous sequence. It is also possible to construct counting sequences in which the successor lists the number of consecutive appearances of a digit: if $C_0 = 2, 1, 1, 4$, then $C_1 = 1, 2, 2, 1, 1, 4$ and $C_2 = 1, 1, 2, 2, 2, 1, 1, 4$. See **[2]** for Conway's analysis of such counting sequences.

2. BASIC PROPERTIES OF COUNTING SEQUENCES. We begin by giving several important properties of the sequences making up a counting sequence, and then give a simple proof of Theorem 1. Fix a finite sequence of positive integers S_0 and let $\{S_i\}_{i \geq 0}$ be the corresponding counting sequence. For $i \geq 1$ we write S_i as

$$S_i = m_{i,1}, f_{i,1}, m_{i,2}, f_{i,2}, \cdots, m_{i,n_i}, f_{i,n_i}.$$

We assume the $f_{i,j}$'s are in increasing order and leave out commas to unclutter the notation when there is no risk of confusion. The positive integer $m_{i,j}$ is called a *multiplier* of S_i and indicates that the integer $f_{i,j}$, called a *factor* of S_i, appears exactly $m_{i,j}$ times in S_{i-1}. Let $|S_i| = 2n_i$ be the total number of terms in S_i. The following observations about the S_i's are used often, and frequently without mention. Similar facts are proved in **[1]** and **[6]**.

Proposition 2. *Fix S_0 and let $\{S_i\}_{i \geq 0}$ be the corresponding counting sequence. Let $i \geq 1$.*

1) *For each factor $f_{i,j}$ of S_i there are $m_{i,j} - 1$ or $m_{i,j}$ multipliers of S_{i-1} with the value $f_{i,j}$, depending on whether or not the value $f_{i,j}$ appears as a factor in S_{i-1}.*
2) *We have $|S_{i-1}| = \sum_{j=1}^{n_i} m_{i,j}$ and $|S_{i-1}| \leq |S_i|$, because every factor of S_{i-1} is also a factor of S_i.*
3) *If $\{S_i\}_{p_0 \leq i \leq p}$ constitutes a p-cycle, then $|S_i| = |S_{i+1}|$ for all i and each S_i in the cycle has exactly the same factors. Further, $|S_{i-1}| = \sum_{j=1}^{n_i}(m_{i,j} - 1)f_{i,j}$.*

To show how these facts will be used we provide the following proof of Theorem 1.

Proof of Theorem 1: Fix S_0, and let $\max(S_i)$ be the value of the largest term in S_i. Clearly either $\max(S_i) = \max(S_2)$ for all $i \geq 2$, or there is some i such that $\max(S_{i+1}) > \max(S_i)$. First assume the former. For $i \geq 2$, the number of sequences S_i with $\max(S_i) \leq n$ for any particular n is at most $(n+1)^n$, and so is finite. Since S_{i+1} is completely determined by S_i, we then see that the counting sequence $\{S_i\}_{i \geq 0}$ must eventually repeat, and so enters a cycle.

So now suppose $\max(S_{i+1}) > \max(S_i)$ for some $i \geq 2$, and choose n so that $n+1$ is a term in S_{i+1} and is larger than every term in S_i. Since $n+1$ can appear in S_{i+1} only as a multiplier, S_i has at least $n+1$ equal terms. But clearly $|S_i| \leq 2n$, and since $i \geq 2$, the factors in S_i are distinct. It must therefore be the case that all of the multipliers of S_i are equal, $|S_i| = 2n$, and each of the integers from 1 to n appears as a multiplier in S_i. Write $S_i = m, 1, m, 2, \ldots, m, n$ for some $m \geq 1$. Then $mn = \sum_{j=1}^{n} m = |S_{i-1}| \leq |S_i| = 2n$ shows $m \leq 2$.

If $m = 2$ then

$$2n \geq |S_{i-1}| \geq \sum_{j=1}^{n} (m-1) f_j = \sum_{j=1}^{n} f_j \geq \sum_{j=1}^{n} j$$

shows that $n \leq 3$, and that S_i must be $2,1$ or $2,1,2,2$ or $2,1,2,2,2,3$. A counting sequence containing any of these is easily shown to converge to $2,1,3,2,2,3,1,4$, a one-cycle. A similar argument shows that if $m = 1$ and $i \geq 2$, then $S_{i-1} = 1,2$ or $1,2,3,4$ or $1,2,3,4,5,6$, all of which lead to periodic counting sequences. ■

3. CYCLES AND THEIR TRUNCATIONS. Theorem 1 ensures that no matter what finite sequence S_0 of positive integers we begin with, the counting sequence associated to S_0 is ultimately periodic, that is, ends in a cycle of some period p. We now determine the possible periods, and for each p classify the p-cycles. As the word "classify" hints, there are actually infinitely many different cycles, and the sequences in these cycles may be arbitrarily long. Fortunately there are only three possible periods, and each cycle has a companion cycle made up of very short sequences. We will use these truncated sequences to make our classification.

Fix a p-cycle, and for ease, rename the sequences in it $S_1, S_2, \ldots, S_p$. We first show that 1 occurs as a term in each S_i, unless the cycle is the one-cycle $S_1 = 2, 2$. This implies that the multiplier of the factor 1 will play an important role in our classification.

Lemma 3. *Either* 1 *occurs at least twice in each* S_i, *or* $p = 1$ *and* $S_1 = 2, 2$.

Proof: First suppose no S_i has 1 as a factor, so all of the multipliers in each S_i have values larger than 1. Let $|S_i| = 2n$. Since the sum of the n multipliers of S_i equals $|S_{i-1}| = |S_i| = 2n$, all of the multipliers of S_i must equal 2. This is true for all i. But then all of the S_i's have exactly the same multipliers, all of value 2, and exactly the same factors, so this cycle is the one-cycle $S_1 = 2, 2$.

Next, when one S_i has 1 as a factor, then each S_i does. If S_{i+1} contains exactly one 1, for some i, then none of S_i's multipliers equal 1. Again, $\sum_{j=1}^{n} m_{i,j} = |S_{i-1}| = |S_i| = 2n$ then implies that all of the multipliers of S_i have the value 2. However, as in the proof of Theorem 1, a counting sequence containing such an element converges to the one-cycle $2,1,3,2,2,3,1,4$, which is not equal to S_{i+1}, contradicting our assumption. Thus each S_i contains at least two 1's, as desired. ■

Next consider the factors whose multipliers are equal to 1. If the factor f of S_i has multiplier 1, then f appears in S_{i-1} only as a factor, so it plays a relatively unimportant role in the creation of S_i. This leads us to consider the *truncation* S_i' of S_i formed by deleting all the multiplier-factor pairs of S_i whose multipliers are 1. For example, if $S_1 = 6, 1, 2, 2, 1, 3, 1, 4, 1, 5, 2, 6, 1, 7$, then $S_1' = 6, 1, 2, 2, 2, 6$. We will see that there are rather few sequences that arise as the truncation of a sequence in a cycle, and so will be able to use truncation to classify the cycles.

Assume S_i is a sequence belonging to a cycle, and by Lemma 3 that S_i' has the form

$$S_i' = m_{i,1}\,1\,m_{i,2}\,f_{i,2}\,\cdots\,m_{i,k_i}\,f_{i,k_i}$$

with $m_{i,j} \geq 2$ for all j. In studying S_i', the first step is to establish a property similar to part 2 of Proposition 2.

Lemma 4. *In a cycle we have* $k_i - 1 = |S_i'|/2 - 1 \leq \sum_{j=2}^{k_i}(m_{i,j} - 1) = |S_{i-1}'|/2$ *for all* i. *In particular,* $|S_i'| \leq |S_{i-1}'| + 2$ *for all* i.

Proof: The first equality and the following inequality are trivial, since $m_{i,j} \geq 2$. For the last equality, since $m_{i,j} - 1$ is the number of multipliers in S_{i-1} with value $f_{i,j}$, the number of multipliers in S_{i-1} that are not equal to 1 is $\sum_{j \geq 2}(m_{i,j} - 1)$. But the multipliers in S_{i-1}' are exactly the multipliers in S_{i-1} that do not equal 1. Thus the sum equals the number of multipliers in S_{i-1}', or $|S_{i-1}'|/2$. ■

Therefore, in a cycle either $|S_i'| = |S_{i-1}'|$ for all i, or $|S_i'| = |S_{i-1}'| + 2$ for some i. Since each multiplier in S_i' is larger than 1, if $|S_i'| = |S_{i-1}'|$ then Lemma 4 shows that $\{m_{i,j} : j \geq 2\}$ consists of all 2's except for possibly one 3, while if $|S_i'| = |S_{i-1}'| + 2$ then $m_{i,j} = 2$ for all $j \geq 2$. We next show that the first case corresponds to the one-cycles, and so the second case corresponds to the longer cycles.

Proposition 5. *Suppose* $\{S_i\}_{1 \leq i \leq p}$ *is a cycle such that* $|S_i'| = |S_{i+1}'|$ *for all* i. *Then* $p = 1$, *so* S_1 *is actually a one-cycle.*

Proof: Since all the S_i's have exactly the same factors, it suffices to show that the multiplier of any particular factor is the same in all of the S_i's. Because the set of multipliers of S_{i-1} is exactly the set of factors of S_i', and we might as well assume S_i is not 2, 2, the only factors of S_i whose multipliers are not 1 are 1, 2, 3, and m, where m is the multiplier of 1 in S_{i-1}. We concentrate on these factors. First, $1 + (|S_i| - |S_i'|)/2$ is independent of i and gives the number of 1's in S_i. Thus $1 + (|S_i| - |S_i'|)/2 = m$, and each of the S_i's contains m 1's. Next, we have seen that each $\{m_{i,j} : j \geq 2\}$ consists of all 2's except for possibly one 3. Since the value of the sum $\sum_{j \geq 2}(m_{i,j} - 1) = |S_{i-1}'|/2$ is independent of i, as m is, we see that each S_i must contain the same number of 2's and the same number of 3's. Finally, for $m \geq 4$, m occurs in each S_i exactly twice, as the multiplier of 1 and as a factor. ■

This proposition allows us to find the truncations of all the one-cycles. Except for $S = 2, 2$, the set of multipliers of a one-cycle consists of 2's, possibly one 3, and the multiplier m of 1. We point out the various cases and let the reader check the details. If $m = 2$ then $S' = 2, 1, 3, 2, 2, 3$, while if $m = 3$ then $S' = 3, 1, 2, 2, 3, 3$ or $S' = 3, 1, 3, 3$ depending on whether 2 is a multiplier of S or not. Since m is at

least 2, the only other possibility is $m \geq 4$, and then $S' = m, 1, 3, 2, 2, 3, 2, m$. We have proved

Theorem 6. *If S is a one-cycle, then S' is $2,2$ or $3,1,3,3$ or $2,1,3,2,2,3$ or $3,1,2,2,3,3$ or $m,1,3,2,2,3,2,m$ for some $4 \leq m$.*

We next turn our attention to the cycles whose periods are longer than 1. From Lemma 4 and Proposition 5 we know that in such a cycle $|S'_i| = |S'_{i-1}| + 2 \geq 4$ for some i, and that the multipliers in S'_i all equal 2, except for possibly the multiplier of 1. In fact, since $2n = |S_i| = |S_{i-1}| = \sum_{j=1}^{n} m_{i,j}$ is a sum of 1's, at least one 2, and the multiplier of 1, we see the multiplier of 1 in S_i must be at least 3. If we write

$$S'_i = m_{i,1}, 1, 2, f_{i,2}, \cdots, 2, f_{i,k_i},$$

for $k_i \geq 2$, then

$$S'_{i+1} = m_{i+1,1}, 1, k_i, 2, 2, m_{i,1}$$

for some m_{i+1}. Thus if $|S'_i| = |S'_{i-1}| + 2$ for some i, then $|S'_{i+1}| = 6$. Because $|S'_i| \leq |S'_{i-1}| + 2$ for all i, it must therefore be the case that $|S'_{i-1}| = 6$, $|S'_i| = 8$, and $|S'_{i+1}| = 6$.

Now write $S'_i = m, 1, 2, a, 2, b, 2, c$ for $m \geq 3$ and some a, b, and c. Clearly S_{i-1} must have $m-1$ multipliers equal to 1. Because $|S'_{i-1}| = 6$, we have $|S_i| = |S_{i-1}| = 2(m+2)$, so the number of 1's in S_i is $1 + (|S_i| - |S'_i|)/2 = (m-1)$. Thus the multiplier of 1 in S_{i+1} is $m-1$. Similarly, the multiplier of 1 in S_{i-1} is m or $m-1$, depending on whether $|S_{i-2}|$ is 6 or 8, so S_i contains either two m's or two $(m-1)$'s.

Before considering these two cases, we show how to construct S'_{i+1} directly from S'_i. Any integer $f \neq 1$ appears as a factor in S'_{i+1} if and only if it appears as a multiplier in S'_i, and then its multiplier in S'_{i+1} is one more than the number of times it appears as a multiplier in S'_i. Lemma 3 shows that $f = 1$ also appear as a factor in S'_{i+1}. The next lemma shows how to compute its multiplier.

Lemma 7. *The number of 1's in S_i is $1 + \sum\big((m_{i,j} - 1)(f_{i,j} - 1) - 1\big)$, where the sum is over the multipliers of S'_i.*

Proof: Let m be the number of 1's appearing as multipliers in S_i. Then

$$\begin{aligned} 1 + &\sum_{m_{i,j} \in S'_i} \big((m_{i,j} - 1)(f_{i,j} - 1) - 1\big) \\ &= 1 + m + \sum_{m_{i,j} \in S_i} \big((m_{i,j} - 1)(f_{i,j} - 1) - 1\big) \\ &= 1 + m + \sum_{m_{i,j} \in S_i} (m_{i,j} - 1) f_{i,j} - \sum_{m_{i,j} \in S_i} m_{i,j}. \end{aligned}$$

The two last sums both equal $|S_{i-1}|$, and $1 + m$ is the number of 1's in S_i. ∎

Consider again $S'_i = m, 1, 2, a, 2, b, 2, c$. If S_i contains two $(m-1)$'s, with $m \geq 3$, then $S'_i = m, 1, 2, a, 2, b, 2, m-1$, for some a and b, and so $S'_{i+1} = m-1, 1, A, 2, 2, m$, for some A. It is then clear that m cannot be 3. By Lemma 7, $a + b = 6$, so $a = 2$ and $b = 4$. Thus $S'_i = m, 1, 2, 2, 2, 4, 2, m-1$ and $m \neq 5$. Using Lemma 7 again we see

$$S'_{i+1} = m - 1, 1, 4, 2, 2, m$$

$$\text{and} \quad S'_{i+2} = m, 1, 2, 2, 2, 4, 2, m - 1 = S'_i,$$

for $m = 4$ or $m \geq 6$, and so this is a two-cycle. Notice that when $m \geq 6$ the multiplier-factor pair $1, m$ must appear in S_i and the pair $1, m-1$ in S_{i+1}.

If S_i contains two m's, then $S_i' = m, 1, 2, a, 2, b, 2, m$ for some a and b, and so $S_{i+1}' = m-1, 1, A, 2, 2, m$ for some A. By Lemma 7, $a + b = 5$, so $a = 2$, $b = 3$ and $S_i' = m, 1, 2, 2, 2, 3, 2, m$. Thus $S_{i+1}' = m-1, 1, 4, 2, 2, m$. If m is not equal to 5, we are led to one of the cycles above. If m does equal 5, then using Lemma 7 we have

$$S_i' = 5, 1, 2, 2, 2, 3, 2, 5$$

$$S_{i+1}' = 4, 1, 4, 2, 2, 5$$

$$S_{i+2}' = 5, 1, 2, 2, 3, 4$$

and $S_{i+3}' = S_i'$, so this is a three-cycle. Notice that the pair $1, 3$ must appear in S_{i+1} and $1, 5$ must appear in S_{i+2}. We have proved

Theorem 8. *Suppose $\{S_i\}_{1 \leq i \leq p}$ is a periodic counting sequence with period $p > 1$. Then either $p = 2$ or $p = 3$. In fact,*

i. *If $p = 2$ then the truncated form of $\{S_i\}$ is* $\begin{cases} S_1' = m, 1, 2, 2, 2, 4, 2, m-1 \\ S_2' = m-1, 1, 4, 2, 2, m \end{cases}$ *with $m = 4$ or $m \geq 6$*

ii. *If $p = 3$ then the truncated form of $\{S_i\}$ is* $\begin{cases} S_1' = 5, 1, 2, 2, 2, 3, 2, 5 \\ S_2' = 4, 1, 4, 2, 2, 5 \\ S_3' = 5, 1, 2, 2, 3, 4. \end{cases}$

It is a simple matter to rebuild a cycle from its truncation just by picking reasonable factors. For instance, if $S' = 3, 1, 2, 2, 3, 3$ then $S = 3, 1, 2, 2, 3, 3, 1, 4, 1, 5$ gives a one-cycle. Of course, so does $S = 3, 1, 2, 2, 3, 3, 1, 5, 1, 20$ and, if we expand our possible choice of factors, so does $S = 1, -4, 1, 0, 3, 1, 2, 2, 3, 3$. In the sequel it will be useful to allow 0 as a factor.

Notice that no three-cycle can contain more than seven factors, or more than two factors larger than 5. Thus, if S_0 contains eight or more distinct numbers, or two or more distinct numbers larger than 5, then the cycle to which its counting sequence converges cannot have period 3. Nor can it converge to any one-cycle, except for one whose truncation has the form $m, 1, 3, 2, 2, 3, 2, m$. Since most finite sequences S_0 contain three different numbers larger than 5, we see most counting sequences converge to a one-cycle of the form $m, 1, 3, 2, 2, 3, 2, m$, or to a two-cycle. In fact, the multiplier of 2 can be used to distinguish between these last two cases, but unfortunately we do not have methods to predict the multipliers of 2. It would be quite interesting to have a more precise answer.

Similarly, we would like to have a method to determine the pre-period of a given S_0, that is, to be able to measure how far S_i is from entering a cycle.

4. CYCLES IN A FINITE BASE. From Theorem 6 the numerical portion of an answer Raphael Robinson's puzzle is

$$1, 0, 1, 7, 1, 3, 2, 2, 3, 1, 4, 1, 5, 1, 6, 2, 7, 1, 8, 1, 9.$$

Is this answer unique? No, there is another:

$$1, 0, 11, 1, 2, 2, 1, 3, 1, 4, 1, 5, 1, 6, 1, 7, 1, 8, 1, 9,$$

if we read 11 as two 1's. This makes sense only if we represent the value *eleven* as $1 \cdot 10^1 + 1 \cdot 10^0$, i.e., if we write our numbers in base 10. This example reveals the

basic and interesting difference between the counting sequences over finite bases and those over the infinite base: when we use a finite base the multipliers can consist of multiple digits, which, by definition, is impossible over the infinite base.

What cycles are possible if we must choose the factors from the digits 0 through $k-1$ and consider the multipliers in base k? If the factors are kept smaller than k, then Theorems 6 and 8 provide examples of cycles in base k. In base 5, for instance,

$$1,0,3,1,1,2,3,3 \quad \text{and} \quad 2,1,3,2,2,3,1,4$$

are one-cycles. However $(11)_5, 1, 1, 2, 1, 3, 1, 4$ is also a one-cycle in base 5, where $(11)_5$ is the representation of the number *six* in base 5, so these theorems do not list all of the cycles. It thus remains for us to find the cycles that contain at least one multiplier with multiple digits in base k.

We first show that given any sequence S_0 the counting sequence $\{S_i\}_{i\geq 0}$ formed in base k is eventually periodic. As in Section 2, it suffices to show that S_i, $i \geq 1$, can take on only finitely many forms. We now write $|S_i|_k$ for the total number of digits appearing in S_i when it is written in base k. For example, $|(11)_3, 1, 1, 2|_3 = 5$.

Lemma 9. *In base $k \geq 4$, we have $|S_i|_k \leq 2k+1$ for all sufficiently large i.*

Proof: We simply show that if $|S_{i-1}|_k \leq |S_i|_k$ for some i, then $|S_i|_k \leq 2k+1$. As in Proposition 2 we have $|S_{i-1}|_k = \sum_{j\geq 1} m_j$, where the m_j's are the multipliers of S_i. Letting $\#m_j$ be the number of digits of m_j in base k, we then have

$$|S_i|_k = \text{the number of factors in } S_i + \sum_{m_j \in S_i} \#m_j \leq k + \sum_{m_j \in S_i} \#m_j. \tag{4.1}$$

Using $\sum m_j = |S_{i-1}|_k \leq |S_i|_k$, we see that

$$\sum_{m_j \in S_i} (m_j - \#m_j) \leq k. \tag{4.2}$$

But $m_j - \#m_j$ is at least $k-2$ if $m_j \geq k \geq 4$. Thus, for $k \geq 5$ there can be at most one multiplier of S_i consisting of multiple digits in base k, and its value can be at most $k+2$. When $k = 4$, one shows easily that a sequence in a counting sequence with two multipliers larger than 3 must have multipliers $4 = (10)_4$, $4 = (10)_4$, 1, and 1, and its counting sequence converges to $1, 0, (11)_4, 1, 1, 2, 2, 1, 3$. Therefore $|S_i|_k \leq 2k+1$, as desired. ■

Thus, when k is at least 4 there are only finitely many sequences that may appear in any given counting sequence. Inequality (4.2), which holds in any base, shows that 5 is the largest possible value of a multiplier in a counting sequence in base 2 or 3, and so also over these two bases a sequence in any given counting sequence may take on only finitely many forms. Therefore, all counting sequences in base k are eventually periodic for all k.

Checking the possibilities, which we leave to the reader, in base 2 the only cycles are $(11)_2, 1$ from Theorem 8, and $(11)_2, 0, (100)_2, 1$. In base 3, Theorem 6 gives only 2, 2, while Theorem 8 gives three one-cycles. The only other one-cycles in base 3 are

$$(10)_3, 0, (10)_3, 1, 2, 2 \quad \text{and} \quad 2, 0, 2, 1, (10)_3, 2 \quad \text{and} \quad (10)_3, 0, (10)_3, 1,$$

and the only longer cycle is $\begin{cases} S_1 = 1, 0, (10)_3, 1, (10)_3, 2 \\ S_2 = (10)_3, 0, (11)_3, 1, 1, 2 \\ S_3 = 2, 0, (12)_3, 1, 1, 2. \end{cases}$

Now suppose k is at least 4, and, by Lemma 9, that S_i is a sequence with one multiplier M such that $k \leq M \leq k+2$. If S_i has f factors, then S_{i-1} has at most f factors, and since they each have no more than one multiplier with two digits, we see that $|S_{i-1}|_k \leq |S_i|_k$. Thus inequality (4.2) may be more accurately stated as

$$k - 2 \leq \sum_{m_j \in S_i} (m_j - \#m_j) \leq (|S_i|_k - 1)/2 \leq k. \tag{4.3}$$

If $M = k+2$, then all of the other multipliers in S_i must equal 1, and $|S_i|_k = 2k+1$. One then sees that S_{i+1} has the form

$$S_{i+1} = 1, 0, (11)_k, 1, 2, 2, 1, 3, \ldots, 1, k\text{-}1, \tag{4.4}$$

which constitutes a one-cycle. If $M = k+1$, then the other multipliers in S_i equal 1, except for possibly one 2. When 2 is a multiplier of S_i, we have $|S_i|_k = 2k+1$ and then S_{i+1} is as given in (4.4). When 2 is not a multiplier in S_i, then $|S_i|_k = 2k-1$ and

$$S_{i+1} = 1, 0, (11)_k, 1, 1, 2, \ldots, \widehat{1, l}, \ldots, 1, k-1$$

for some $0 \leq l \leq k-1$, $l \neq 1$, where $\widehat{1, l}$ means that this pair does not appear in S_{i+1}. Clearly S_{i+1} forms a one-cycle. Finally, if $M = k = (10)_k$, then it is not difficult to use part 2) of Proposition 2 to show that $\{S_i\}_{i \geq 0}$ converges to a one-cycle consisting of terms having one base k digit each. We have proved

Proposition 10. *The only cycles in base $k \geq 4$ that have multipliers with two or more digits are one-cycles. Further, if S' is the truncated version of one of these sequences, then S' is either $(11)_k, 1$ or $(11)_k, 1, 2, 2$.* ■

We have now discovered all possible cycles in base k. Since k is finite, the number of cycles is finite and can be counted.

Theorem 11. *For $k \geq 4$, the number of one-cycles in base k is $2^{k-4} + k(k-1)/2$. In bases 4 and 5 there are no longer cycles, while in base $k \geq 6$ there are $2^{k-5} - 1$ two-cycles, $\binom{k-5}{2}$ three-cycles, and no longer cycles.*

The proof of Theorem 11 is just a matter of undoing the truncation process, and then using the binomial theorem. For example, for each $m \geq 4$ and $k \geq 6$ there are $\binom{k-4}{m-1}$ one-cycles S with $S' = m, 1, 3, 2, 2, 3, 2, m$, so there are

$$\sum_{m=4}^{k-1} \binom{k-4}{m-1} = 2^{k-4} - 1 - (k-4) - \binom{k-4}{2}$$

one-cycles S with S' having the form $m, 1, 3, 2, 2, 3, 2, m$. We leave the rest of the proof to the reader.

5. INFINITE SEQUENCES AND INFINITE CYCLES. From Theorem 1 we know that every counting sequence beginning with a finite sequence S_0 is ultimately periodic. Is this true when S_0 is an infinite sequence? In this section we show that it is not and provide two methods for constructing counter-examples.

If one chooses an infinite sequence S_0 at random, its associated counting sequence may fail to exist. For example, if we choose $S_0 = 1, 2, 3, 4, 5, 6, \ldots$, then $S_1 = 1, 1, 1, 2, 1, 3, 1, 4, 1, 5, 1, 6, \ldots$, but S_2 is not well-defined and so $\{S_i\}_{i \geq 0}$ does

not exist. It would be interesting to have necessary or sufficient conditions on S_0 so that its counting sequence exists. We will not concern ourselves here with general existence and convergence questions but instead concentrate on supplying a variety of examples.

We begin by constructing infinite sequences whose associated counting sequences are actually one-cycles. First, let $S_0^0 = 4, 4$, and define $S_0^1 = 4, 4, 4, 5, 4, 6$. Notice that there are four 4's in S_0^1, which fits the description given in S_0^0. Next, create S_0^2 to fit the description given in S_0^1 and to have consecutive factors, and then similarly create S_0^3 by the description implicit in S_0^2:

$$S_0^2 = 4,4,4,5,4,6,5,7,5,8,5,9,6,10,6,11,6,12$$

$$\begin{aligned} S_0^3 = {} & 4,4,4,5,4,6,5,7,5,8,5,9,6,10,6,11,6,12,7,13,7,14,7,15,7,16,8,17,8, \\ & 18,8,19,8,20,9,21,9,22,9,23,9,24,10,25,10,26,10,27,10,28,10,29,11, \\ & 30,11,31,11,32,11,33,11,34,12,35,12,36,12,37,12,38,12,39. \end{aligned}$$

Finally, define S_0 to be the limit of the finite sequences $\{S_0^k\}$. It is then clear that S_0 forms a one-cycle, and so each element of the counting sequence $\{S_i\}$ exists. We adopt the terminology of [3] to call the process that takes the finite sequence S_0^0 and produces the infinite sequence S_0 the *self-generating process*.

Proposition 12. *Let $S = m_1, f_1, m_2, f_2, \ldots, m_n, f_n$ be a sequence of positive integers such that the f_i are strictly increasing, f_i appears no more than m_i times in S, and each m_i also appears as an f_j. Then, setting $S_0^0 = S$, the sequences S_0^k can be constructed using the self-generating process, $S_0 = \lim_k S_0^k$ exists, and S_0 forms a one-cycle.*

To give the relative sizes of the factors and multipliers of our particular example $S_0 = 4, 4, 4, 5, 4, 6, \ldots$ we introduce an integer sequence constructed and studied first by Golomb [3]. This sequence

$$1, 2, 2, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 8, \cdots$$

consists of the values of the function $G(n)$ defined on the natural numbers by

(i) $G(1) = 1$
(ii) $G(n) = \#\{\text{integers } m : G(m) = n\}$
(iii) $G(n)$ is non-decreasing.

Golomb proved the asymptotic formula $G(n) \sim \phi(n/\phi)^{\phi-1}$, where $\phi = (\sqrt{5} + 1)/2$ is the golden ratio. If we replace the first three terms of Golomb's sequence by a 3, and then add 1 to each term, the resulting sequence consists of the multipliers of S_0. Thus, the multiplier of f in S_0 is approximately $G(f)$. Inverting the asymptotic formula for $G(f)$ then gives

Proposition 13. *Let m_f be the multiplier of f in $S_0 = 4, 4, 4, 5, \ldots$. Then*

$$f \sim \phi\left(\frac{m_f}{\phi}\right)^{\phi}.$$

It is a simple matter to modify S_0 to create counting sequences consisting of infinite sequences that converge to longer cycles. For instance, define $S_0(24)$ to be the sequence that is identical to S_0 except that the multiplier-factor pair 9, 24, is replaced by 10, 24, i.e.,

$$\begin{aligned} S_0(24) = {} & 4,4,4,5,4,6,5,7,5,8,5,9,6,10,6,11,6,12,7,13,7,14,7,15,7,16,8,17, \\ & 8,18,8,19,8,20,9,21,9,22,9,23,\underline{10,24},10,25,10,26,10,27,10,28,\ldots. \end{aligned}$$

We have underlined the multiplier-factor pairs of $S_0(24)$ that do not agree exactly with S_0, i.e., the positions of $S_0(24)$ that are in "error" when compared to S_0. If $S_1(24)$ is the usual description of the sequence $S_0(24)$, then $S_1(24)$ contains one more 10 but one fewer 9 than S_0 contains, so

$$\begin{aligned} S_1(24) = {} & 4,4,4,5,4,6,5,7,5,8,\underline{4,9,7,10},6,11,6,12,7,13,7,14,7,15,7,16,8,17, \\ & 8,18,8,19,8,20,9,21,9,22,9,23,9,24,10,25,10,26,10,27,10,28,\cdots, \end{aligned}$$

and

$$\begin{aligned} S_2(24) = {} & \underline{5,4,3,5,3,6,6,7},5,8,5,9,6,10,6,11,6,12,7,13,7,14,7,15,7,16,8,17, \\ & 8,18,8,19,8,20,9,21,9,22,9,23,9,24,10,25,10,26,10,27,10,28,\cdots. \end{aligned}$$

Notice that the multiplier-factor pair in error in $S_0(24)$ has been "repaired" in $S_1(24)$, and that the errors in $S_1(24)$ are repaired in $S_2(24)$. Also notice that the numerical values of the multipliers in error in S_0 and factors in error in S_1 are very close. The same is true for the multipliers in error in S_1 and the factors in error in S_2. If we continue, the counting sequence converges to

$$\begin{aligned} S_{10}(24) &= \underline{1,1,4,2,2,3,2,4,5,5},4,6,5,7,5,8,5,9,6,10,\cdots \\ S_{11}(24) &= \underline{2,1,3,2,1,3,3,4,5,5},4,6,5,7,5,8,5,9,6,10,\cdots \\ S_{12}(24) &= \underline{2,1,2,2,3,3,2,4,5,5},4,6,5,7,5,8,5,9,6,10,\cdots \\ S_{13}(24) = S_{10}(24) &= \underline{1,1,4,2,2,3,2,4,5,5},4,6,5,7,5,8,5,9,6,10,\cdots. \end{aligned}$$

So we have constructed an example of an infinite three-cycle.

We can abstract two facts from this example. Suppose we create an infinite sequence $S_0(f)$ that is identical to an infinite one-cycle S_0 except that the multiplier of f in S_0 has been increased by one. Then, (1), at the beginning of the counting sequence $\{S_i(f)\}_{i \geq 0}$ the errors move quickly to the "left", and (2) once the errors have reached the beginning of the sequences, they (relatively) quickly settle into a cycle. It is not too difficult to convince oneself of these facts, because Proposition 13 tells us that a multiplier in S_0 is far smaller than its factor.

More generally, one may define $S_0(f_1, f_2, \ldots, f_m)$ to be identical to an infinite one cycle S_0 except that the multipliers of the f_j's in S_0 have been increased by one, and then consider the counting sequence $\{S_i(f_1, f_2, \ldots, f_m)\}_{i \geq 0}$. Each such counting sequence ends in a cycle. It would be interesting to classify the cycles that arise in this manner.

Using these ideas we can describe the construction of an infinite counting sequence that is *not* ultimately periodic. For a given integer f_j let n_j be the pre-period of $\{S_i(f_j)\}_{i \geq 0}$. That is, $S_i(f_j)$ is part of a cycle if $i \geq n_j$. Choose an infinite set of factors f_j, $j \geq 1$, growing fast enough in j so that for all $i \leq n_j$ and $f < f_{j-1}$, the multipliers of the factor f in S_0 and $S_i(f_j)$ are equal. In other words, choose f_j so that it takes more than n_j steps for the errors in $S_i(f_j)$ to move themselves to the point of the initial error in $S_0(f_{j-1})$. Once we fix such an infinite sequence, then $\{S_i(f_1, f_2, f_3, \cdots)\}_{i \geq 0}$ will be a non-periodic counting sequence. To actually construct such a family of f_j's one needs to use Proposition 13 to give a careful study of the rates at which the errors in $\{S_i(f_j)\}$ spread and move to the left.

As this study would occupy the better part of several pages, we instead end this section with a very simple method for constructing infinite counting sequences that

are both well-defined and not ultimately periodic. Define $\{S_i\}_{i\geq 1}$ by the following rules:

(i) The multiplier of i in S_i has value at least i.
(ii) Every natural number occurs as a factor in each S_i.
(iii) The multipliers in each S_i form a non-decreasing sequence.
(iv) S_{i+1} is the description of S_i for $i \geq 1$.

Rule (i) insures that the terms below the main diagonal are not influenced by those above the main diagonal. For instance, taking the multiplier of i to be $i+1$ gives the following:

$$S_1 = \mathbf{2,1},2,2,3,3,3,4,4,5,4,6,5,7,5,8,5,9,6,10,6,11,6,12,7,13,7,14,7,15,\cdots$$
$$S_2 = 1,2,\mathbf{3,2},3,3,3,4,4,5,4,6,4,7,5,8,5,9,5,10,6,11,6,12,6,13,7,14,7,15,\cdots$$
$$S_3 = 1,1,2,2,\mathbf{4,3},4,4,4,5,4,6,5,7,5,8,5,9,5,10,6,11,6,12,6,13,6,14,7,15,\cdots$$
$$S_4 = 2,1,2,2,1,3,\mathbf{5,4},5,5,5,6,5,7,5,8,6,9,6,10,6,11,6,12,6,13,7,14,7,15,\cdots$$
$$S_5 = 2,1,3,2,1,3,1,4,\mathbf{6,5},6,6,6,7,6,8,6,9,6,10,7,11,7,12,7,13,7,14,7,15,\cdots$$
$$S_6 = 3,1,2,2,2,3,1,4,1,5,\mathbf{7,6},7,7,7,8,7,9,7,10,7,11,7,12,8,13,8,14,8,15,\cdots$$
$$S_7 = 3,1,3,2,2,3,1,4,1,5,1,6,\mathbf{8,7},8,8,8,9,8,10,8,11,8,12,8,13,8,14,9,15,\cdots$$
$$S_8 = 4,1,2,2,3,3,1,4,1,5,1,6,1,7,\mathbf{9,8},9,9,9,10,9,11,9,12,9,13,9,14,9,15,\cdots.$$

In S_{i+1} the multiplier of i is either 2 or 1, depending on whether the multiplier of i in S_i is i or greater than i. Therefore $\{S_i\}_{i\geq 1}$ is well-defined but is not ultimately periodic.

6. FACTOR-FREE COUNTING SEQUENCES. We end this paper the way we began it: by using the sequence $2,1,1,4$ to build a type of counting sequence. Because $2,1,1,4$ consists of 2 ones, 1 two, 0 threes, and 1 four, let us define R_1 to be the numbers making up this description: $R_1 = 2,1,0,1$. Repeating this process, R_1 consists of 1 zero, 3 ones, 1 two, 0 threes, and 0 fours, so set $R_2 = 1,2,1,0,0$. Continuing we have

$$R_3 = 2,2,1,0,0$$
$$R_4 = 2,1,2,0,0$$
$$R_5 = 2,1,2,0,0.$$

We call the sequence $\{R_i\}_{i\geq 1}$ a *factor-free* counting sequence. The cycles of factor-free sequences are called *self-descriptive* and *co-descriptive strings* in **[5]**, **[7]**, and **[8]**.

Since a factor-free counting sequence is built without the explicit benefit of place-keeping factors, we need a method for indicating which integer each term in each R_i describes. For $i \geq 1$ we assume that the j-th entry of R_i gives the number of times $j-1$ appears in R_{i-1}, and that this entry is 0 if $j-1$ does not appear in R_{i-1} but some integer at least as large as $j-1$ appears in some $R_{i'}$, $1 \leq i' < i$. Then, just as the first number in an element S_i of a counting sequence almost always describes the number of 1's in S_{i-1}, the first number in an element R_i of a factor-free counting sequence describes the number of 0's in R_{i-1}.

Of course, one may allow the first digit of a sequence to describe numbers other than 0. The only finite example of this is the one-cycle 1, which is the factor-free version of the one-cycle 2, 2. There are, however, many infinite examples. For instance, Golomb's sequence can be thought of as an infinite factor-free one-cycle that begins by describing the number of 1's it contains.

Using techniques similar to those in Section 2 it is easy to show that if R_0 is a finite sequence of non-negative integers, then the factor-free counting sequence $\{R_i\}_{i \geq 0}$ is ultimately periodic. To find all of the possible cycles we will relate the factor-free and "ordinary" counting sequences. Following **[1]**, say that an element S_i of a counting sequence is *complete* if its factors are consecutive and the smallest factor is 1. If $S = m_1, f_1, m_2, f_2, \ldots, m_k, f_k$ is a complete element of a counting sequence, then defining $R = n_0, n_1, \ldots, n_k$ by $n_j = m_{j+1} - 1$ gives a factor-free sequence. Similarly, given a sequence $R = n_0, n_1, \ldots, n_k$ of a factor-free counting sequence, defining S by $f_j = j + 1$ and $m_j = n_{j-1} + 1$ gives a factor-containing sequence. Notice that the sequence S corresponding to R is complete. While it is not true that this process allows one to convert between counting sequences $\{S_i\}_{i \geq 1}$ and factor-free counting sequences $\{R_i\}_{i \geq 1}$, it is very easy to show that there is a one-to-one correspondence between the cycles of factor-free counting sequences and the cycles of complete counting sequences. Since there is also a one-to-one correspondence between complete cycles and the truncations appearing in Section 3, Theorems 6 and 8 give our final result.

Corollary 18. *Other than* 1, *the cycles of factor-free counting sequences all contain zeros, and have length one, two, or three. The one-cycles are* $2,0,2,0$ *and* $1,2,1,0$ *and* $2,1,2,0,0$ *and* $m+3,2,1,(m0\text{'s}),1,0,0,0$ *for* $m \geq 0$. *The two-cycles are*

$$\begin{cases} 3,1,1,1,0,0 \\ 2,3,0,1,0,0 \end{cases} \quad \textit{and} \quad \begin{cases} m+3,1,0,1,(m0\text{'s}),1,0 \\ m+2,3,0,0,(m0\text{'s}),0,1, \end{cases}$$

for $m \geq 2$. *Finally, the only other cycle of any length is the three-cycle*

$$\begin{cases} 4,1,1,0,1,0,0 \\ 3,3,0,0,1,0,0 \\ 4,1,0,2,0,0,0. \end{cases}$$

REFERENCES

1. Victor Bronstein and Aviezri S. Fraenkel, On a Curious Property of Counting Sequences, *Amer. Math. Monthly* **101** (1994), 560–563.
2. J. H. Conway, The Weird and Wonderful Chemistry of Audioactive Decay, *Open Problems in Communication and Computation*, T. M. Cover and B. Gopinath, eds., Springer-Verlag, New York, 1987, pp. 173–188.
3. S. Golomb, Problem 5407, *Amer. Math. Monthly* **73** (1966), 674.
4. Douglas Hopfstadter, *Metamagical Themas*, Basic Books, New York, 1985, p. 392.
5. Steven Kahan, A Curious Sequence, *Math. Mag.* **48** (1975), 290–292.
6. Hervé Lehning, Computer-Aided or Analytic Proof?, *College Math. J.* **21** (1990), 228–239.
7. Michael D. McKay and Michael S. Waterman, Self-descriptive Strings, *Math. Gazette* **66** (1982), 1–4.
8. Lee Sallows and Victor L. Eijkhout, Co-descriptive Strings, *Math. Gazette* **70** (1986), 1–10.
9. N. J. A. Sloane and Simon Plouffe, *The Encyclopedia of Integer Sequences*, Academic Press, London, 1995.
10. Marilyn Vos Savant, Ask Marilyn, *Parade Magazine*, May 5th (1996), 6.

Department of Mathematical Science
St. Mary's College
Moraga, CA 94575
jsauerbe@stmarys-ca.edu

Department of Mathematics and Statistics
University of Vermont
Burlington, VT 05401
lshu@math.uvm.edu

Characterizing Continuity

Daniel J. Velleman

It is well known that for any function $f\colon \mathbf{R} \to \mathbf{R}$, f is continuous in the ε-δ sense if and only if the inverse image under f of every open set is open. Recently, when I was teaching a topology class, a student[1] asked if continuity could be characterized using images instead of inverse images. Of course, we cannot simply replace "inverse image" with "image" in the preceding characterization, but perhaps the characterization would work if we used some family of sets other than the family of open sets. Thus, we are led to consider the following question: Does there exist a family $\mathscr{F}$ of sets of reals such that for every function $f\colon \mathbf{R} \to \mathbf{R}$, f is continuous if and only if for every $X \in \mathscr{F}$, $f(X) \in \mathscr{F}$?

My initial reaction was that the answer was probably "no," but I found it surprisingly difficult to prove the "no" answer. Perhaps the reason that the "no" answer is so difficult to prove is that the answer is almost "yes." More precisely, we have the following characterizations of continuity:

Theorem 1. *There is a family $\mathscr{F}$ of sequences of real numbers such that for every function $f\colon \mathbf{R} \to \mathbf{R}$, f is continuous if and only if for every sequence $\{x_n\}_{n=1}^{\infty} \in \mathscr{F}$, $\{f(x_n)\}_{n=1}^{\infty} \in \mathscr{F}$.*

Theorem 2. *There are families $\mathscr{F}$ and $\mathscr{G}$ of sets of reals such that for every function f: $\mathbf{R} \to \mathbf{R}$, f is continuous if and only if for every $X \in \mathscr{F}$, $f(X) \in \mathscr{F}$ and for every $X \in \mathscr{G}$, $f(X) \in \mathscr{G}$.*

Proof of Theorem 1: Let $\mathscr{F}$ be the set of all convergent sequences. If f is continuous and $\{x_n\}_{n=1}^{\infty}$ converges to x, then $\{f(x_n)\}_{n=1}^{\infty}$ converges to $f(x)$. For the converse, suppose f is discontinuous at x. Then there is some sequence $\{y_n\}_{n=1}^{\infty}$ converging to x such that $\{f(y_n)\}_{n=1}^{\infty}$ does not converge to $f(x)$. Define a sequence $\{x_n\}_{n=1}^{\infty}$ letting $x_{2n} = y_n$ and $x_{2n-1} = x$. Then $\{x_n\}_{n=1}^{\infty}$ converges to x, but $\{f(x_n)\}_{n=1}^{\infty}$ does not converge. ■

Proof of Theorem 2: Let $\mathscr{F}$ be the set of all connected sets (i.e., the set of all open, closed, and half-open intervals), and let $\mathscr{G}$ be the set of all compact sets. Clearly both $\mathscr{F}$ and $\mathscr{G}$ are closed under continuous images. Now suppose f is not continuous, and the image under f of every connected set is connected. Choose numbers x and $\varepsilon > 0$ such that for every $\delta > 0$ there is some y such that $|x - y| < \delta$ but $|f(x) - f(y)| \geq \varepsilon$. For each positive integer n choose a number y_n such that $|x - y_n| < 1/n$ but $|f(x) - f(y_n)| \geq \varepsilon$. Then either $f(y_n) \geq f(x) + \varepsilon$ or $f(y_n) \leq f(x) - \varepsilon$. In the first case, since the image under f of the interval from x to y_n is connected, there must be some x_n between x and y_n such that $f(x_n) =$

[1]Actually, the "student" was Amherst College Philosophy professor Alexander George, who was auditing the class. I would like to thank him for suggesting this stimulating problem.

$f(x) + \varepsilon(1/2 + 1/(n+1))$. Similarly, if $f(y_n) \le f(x) - \varepsilon$ then we can choose x_n between x and y_n such that $f(x_n) = f(x) - \varepsilon(1/2 + 1/(n+1))$. Now let $X = \{x_n | n \in \mathbf{Z}^+\} \cup \{x\}$. Then since $\{x_n\}_{n=1}^{\infty}$ converges to x, the set X is closed and bounded, and therefore compact. But either $f(x) + \varepsilon/2$ or $f(x) - \varepsilon/2$ is a limit point of $f(X)$ that is not an element of $f(X)$, so $f(X)$ is not closed and therefore not compact. ■

Theorem 1 can be generalized to any metric space (see [**3**, Theorem 9.3.1]) or even any first countable Hausdorff space. Theorem 2 seems harder to generalize. The proof we have given generalizes to $\mathbf{R}^n$, but not, for example, to the set $\mathbf{Q}$ of rational numbers with the topology induced by the usual metric. To see why, define $f: \mathbf{Q} \to \mathbf{Q}$ by letting $f(x) = 0$ if $x < 0$ and $f(x) = 1$ if $x \ge 0$. Since the only nonempty connected sets in $\mathbf{Q}$ are singletons, the image under f of every connected set is connected, and the image of every set is finite and therefore compact. However, f is not continuous.

Despite Theorems 1 and 2, it turns out that the answer to our original question is indeed "no." In fact, we have the following slightly stronger theorem:

Theorem 3. *There do not exist families of sets of reals $\mathscr{F}$ and $\mathscr{G}$ such that for every function $f: \mathbf{R} \to \mathbf{R}$, f is continuous if and only if for every $X \in \mathscr{F}$, $f(X) \in \mathscr{G}$.*

Proof: Suppose there were families $\mathscr{F}$ and $\mathscr{G}$ as in the theorem. We will draw several conclusions about $\mathscr{F}$ and $\mathscr{G}$ that will eventually lead to a contradiction. Note first that if $\varnothing \in \mathscr{F}$ then since $f(\varnothing) = \varnothing$ for every function f, $\varnothing \in \mathscr{G}$. It follows that if f is discontinuous then there must be a *nonempty* set $X \in \mathscr{F}$ such that $f(X) \notin \mathscr{G}$. In particular, $\mathscr{F}$ must contain at least one nonempty set.

Claim 1. The family $\mathscr{G}$ contains all singletons.

Proof: Let X be any nonempty element of $\mathscr{F}$. Then all continuous images of X must be in $\mathscr{G}$. Since all constant functions are continuous, this means that $\mathscr{G}$ contains all singletons.

Claim 2. The family $\mathscr{G}$ contains no two-element sets.

Proof: Consider any two distinct real numbers, a and b. Let $f: \mathbf{R} \to \mathbf{R}$ be defined as follows:

$$f(x) = \begin{cases} a & \text{if } x < 0 \\ b & \text{if } x \ge 0. \end{cases}$$

Then f is discontinuous, so there must be some nonempty set $X \in \mathscr{F}$ such that $f(X) \notin \mathscr{G}$. But by Claim 1 $f(X)$ can't be a singleton, so the only other possibility is that $f(X) = \{a, b\}$ and therefore $\{a, b\} \notin \mathscr{G}$.

Note that the set X in the proof of Claim 2 must have more than one element, since $f(X) = \{a, b\}$. Since all continuous images of X are in $\mathscr{G}$, it follows that $\mathscr{G}$ must contain some sets with more than one element.

Claim 3. If $X \in \mathscr{F}$, $a, b \in X$, and $a \le c < d \le b$, then $X \cap (c, d) \ne \varnothing$.

Proof: Define a function $f\colon \mathbf{R} \to \mathbf{R}$ as follows:

$$f(x) = \begin{cases} 0 & \text{if } x \le c \\ \dfrac{x-c}{d-c} & \text{if } c < x < d \\ 1 & \text{if } x \ge d. \end{cases}$$

Then f is continuous, so $f(X) \in \mathscr{G}$. But if $X \cap (c, d) = \varnothing$ then $f(X) = \{0, 1\}$, and by Claim 2 this cannot be an element of $\mathscr{G}$. Therefore $X \cap (c, d) \neq \varnothing$.

In the rest of the proof we will use sets related to the Cantor set. Recall that to construct the Cantor set we begin with the closed interval $[0, 1]$ and remove the open middle third $(1/3, 2/3)$. Then we remove the open middle thirds from the two remaining intervals, and then remove the open middle thirds of the remaining intervals, and so on. The Cantor set is what remains after infinitely many such steps. Alternatively, the Cantor set can be described as the set of all numbers in the interval $[0, 1]$ that can be written in base 3 using only the digits 0 and 2.

Claim 4. If $Y \in \mathscr{G}$ and Y has more than one element then Y is uncountable.

Proof: Suppose $Y \in \mathscr{G}$, Y has more than one element, but Y is countable. Enumerating the elements of Y, we can write $Y = \{y_1, y_2, y_3, \dots\}$, where the y_n's need not all be distinct, but they are not all identical.

We now carry out a construction similar to the construction of the Cantor set, except that at each stage we remove *closed* middle thirds. We begin with the open interval $(0, 1)$ and remove the closed middle third $A_1 = [1/3, 2/3]$, leaving the set $B_1 = (0, 1/3) \cup (2/3, 1)$. Then we remove the closed middle third set $A_2 = [1/9, 2/9] \cup [7/9, 8/9]$ from B_1, leaving $B_2 = (0, 1/9) \cup (2/9, 1/3) \cup (2/3, 7/9) \cup (8/9, 1)$. In general, for every n, B_n will be a union of 2^n disjoint open intervals, each having length $1/3^n$. We let A_{n+1} be the union of the closed middle thirds of all of these intervals, and we let $B_{n+1} = B_n \setminus A_{n+1}$.

Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of positive integers such that every positive integer occurs infinitely many times in the sequences. For example, we could use the sequence $1, 1, 2, 1, 2, 3, 1, 2, 3, 4, \dots$. Now define a function $f\colon \mathbf{R} \to \mathbf{R}$ as follows: If $x \in A_n$ for some n, then let $f(x) = y_{a_n}$. Otherwise let $f(x) = y_1$. Clearly f is discontinuous, since the range of f is Y, which is a disconnected set. Thus, there must be some nonempty set $X \in \mathscr{F}$ such that $f(X) \notin \mathscr{G}$. We will show that $f(X) = Y$, which will contradict the fact that $Y \in \mathscr{G}$.

By Claim 1, $f(X)$ is not a singleton, so we can choose $x_1, x_2 \in X$ such that $f(x_1) \neq f(x_2)$. Thus $f(x_1)$ and $f(x_2)$ are not both equal to y_1, so we may assume without loss of generality that $f(x_1) \neq y_1$ and therefore $x_1 \in A_m$ for some m. By definition, A_m is a union of disjoint closed intervals, and x_1 is in one of these intervals. Let $[c, d]$ be the component of A_m containing x_1. Since $f(x_2) \neq f(x_1)$, $x_2 \notin [c, d]$, so either $x_2 < c$ or $x_2 > d$. The two cases are very similar, so we consider only the case $x_2 < c$.

Choose $n \ge m$ large enough so that $x_2 \le c - 1/3^n$. Then B_n is made up of 2^n open intervals, one of which is $(c - 1/3^n, c)$, and $x_2 \le c - 1/3^n < c \le x_1$. It is clear from the construction that for every $t > n$, at least one of the intervals making up A_t is contained in the interval $(c - 1/3^n, c)$. Therefore, by Claim 3, X contains an element of A_t, so $f(X)$ contains y_{a_t}. Since this is true for every $t > n$, and every positive integer occurs infinitely many times in the sequence $\{a_n\}_{n=1}^{\infty}$, it follows that $f(X) = Y$. This completes the proof of Claim 4.

The last step of the proof is motivated by the fact that there is a continuous function, the Cantor-Lebesgue function, that takes on a different constant value on each of the countably many intervals making up the complement of the Cantor set; see [**2**, pp. 37–38]. It follows that any element of $\mathscr{F}$ that is not completely contained in one of these intervals must contain at least one element of the Cantor set, since otherwise its image would be an element of $\mathscr{G}$ that is countable but not a singleton, contradicting Claim 4. We will apply similar reasoning to a family of sets related to the Cantor set.

To define the sets we will be interested in it will be convenient to use base-3 notation. We will say that a function mapping the positive integers to the set $\{0,1,2\}$ is a *digit sequence*. If m is an integer and f is a digit sequence, let $x_{m,f} = m + \sum_{i=1}^{\infty} f(i)/3^i$. Note that if f and g are digit sequences, $f(i) = g(i)$ for all $i < j$, and $f(j) < g(j)$, then $x_{m,f} \leq x_{m,g}$, with equality holding if and only if $f(i) = 2$ for all $i > j$, $g(i) = 0$ for all $i > j$, and $g(j) = f(j) + 1$.

Suppose A is a set of positive integers that is both infinite and co-infinite (i.e., has infinite complement). Motivated by the fact that the Cantor set contains those numbers that can be written in base 3 using only the digits 0 and 2, we define a digit sequence f to be *Cantor-like on A beyond N* if for every $n > N$, $f(n) = 1$ if and only if $n \notin A$. We will say that f is *eventually Cantor-like on A* if it is Cantor-like on A beyond N for some positive integer N. Let

$$X_A = \{x_{m,f} | m \text{ is an integer and } f \text{ is a digit sequence that is eventually Cantor-like on } A\}.$$

Note that, although some numbers have more than one base-3 expansion, the fact that A is co-infinite guarantees that every element of X_A has infinitely many 1's in its base-3 expansion, and therefore that this expansion is unique. In other words, if $x \in X_A$ then there is a unique integer m and a unique digit sequence f such that $x = x_{m,f}$. Note also that X_A is dense in $\mathbf{R}$, and if B is another infinite, co-infinite set of positive integers such that $A \setminus B$ is also infinite, then X_A and X_B are disjoint.

Claim 5. For every infinite, co-infinite set of positive integers A and every $X \in \mathscr{F}$, if X has more than one element that $X \cap X_A \neq \varnothing$.

Proof: Suppose that A is an infinite, co-infinite set of positive integers and $X \in \mathscr{F}$, and suppose also that $c, d \in X$ and $c < d$. Choose an integer m, a positive integer N, and a function g: $\{1, 2, \dots, N\} \to \{0, 1, 2\}$ such that

$$c < m + \sum_{i=1}^{N} \frac{g(i)}{3^i} < m + \sum_{i=1}^{N} \frac{g(i)}{3^i} + \frac{1}{3^N} < d,$$

and note that for every digit sequence f extending $g, c < x_{m,f} < d$. We will say that a digit sequence f is *good* if f extends g and f is Cantor-like on A beyond N. Let

$$W = \{x_{m,f} | f \text{ is a good digit sequence}\}.$$

Then $W \subseteq X_A$ and $W \subseteq (c,d)$. We will show that $X \cap W \neq \varnothing$.

Let $A \setminus \{1, 2, \dots, N\} = \{a_1, a_2, a_3, \dots\}$, with $a_1 < a_2 < a_3 < \cdots$. Imitating the definition of the Cantor-Lebesgue function, we now define a function F: $W \to [0,1]$ as follows: If $x \in W$ then $x = x_{m,f}$ for some unique good digit sequence f. Note that f is determined by its values on the a_i's, and for every i, $f(a_i)$ is either 0 or 2. Let $F(x) = \sum_{i=1}^{\infty} (f(a_i)/2)/2^i$. Then F is a nondecreasing function whose range is the entire interval $[0,1]$.

Now define $\bar{F}: \mathbf{R} \to [0,1]$ by the formula

$$\bar{F}(x) = \sup(\{0\} \cup \{F(y) | y \in W, y \leq x\}).$$

Then $\bar{F}$ extends F and $\bar{F}$ is a nondecreasing function mapping $\mathbf{R}$ onto $[0,1]$, so it is continuous. We claim that for every $x \notin W$, $\bar{F}(x)$ is rational. To see why, consider any $x \notin W$. Since F maps W onto $[0,1]$, we can choose $y \in W$ such that $F(y) = \bar{F}(x)$. By the definition of $W, y = x_{m,f}$ for some good digit sequence f. Since $y \in W$ and $x \notin W$, $y \neq x$, so either $y < x$ or $y > x$. Suppose that $y < x$, and choose n large enough that $y + 2/3^{a_n} \leq x$. If there is some $i > n$ such that $f(a_i) = 0$, then let g be the same as f except that $g(a_i) = 2$, and let $z = x_{m,g}$. Then $z = y + 2/3^{a_i} < y + 2/3^{a_n} \leq x$, $z \in W$, and $F(z) = F(y) + 1/2^i > F(y) = \bar{F}(x)$, contradicting the definition of $\bar{F}(x)$. Therefore for every $i > n$, $f(a_i) = 2$. But then

$$\bar{F}(x) = F(y) = \sum_{i=1}^{\infty} \frac{f(a_i)/2}{2^i} = \sum_{i=1}^{n} \frac{f(a_i)/2}{2^i} + \sum_{i=n+1}^{\infty} \frac{1}{2^i} = \sum_{i=1}^{n} \frac{f(a_i)/2}{2^i} + \frac{1}{2^n}$$

which is rational, as required. A similar argument shows that $\bar{F}(x)$ is rational if $y > x$.

Since $\bar{F}$ is continuous and $X \in \mathscr{F}$, it follows that $\bar{F}(X) \in \mathscr{G}$. Recall that we have $c, d \in X$ and $W \subseteq (c,d)$, so $\bar{F}(c) = 0$ and $\bar{F}(d) = 1$. Therefore $0, 1 \in \bar{F}(X)$, so $\bar{F}(X)$ is uncountable by Claim 4, and therefore $\bar{F}(x)$ is irrational for some $x \in X$. But as we have already observed, $\bar{F}$ maps every number not in W to a rational number, so $x \in W$. Thus $x \in X \cap W$, so $X \cap W \neq \varnothing$. This completes the proof of Claim 5.

Fix a bijection h from the rationals to the positive integers. For every real number r, let $A_r = \{h(q) | q$ is a rational number and $q < r\}$, an infinite, co-infinite set of positive integers. If $r < s$ then, since there are infinitely many rational numbers between r and s, $A_s \setminus A_r$ has infinitely many elements. It follows that X_{A_r} and X_{A_s} are disjoint.

Choose a set $Y \in \mathscr{G}$ with more than one element, and let y_0 be any element of Y. Define a function $f: \mathbf{R} \to \mathbf{R}$ as follows: If $x \in X_{A_r}$ for some $r \in Y$, then let $f(x) = r$. Otherwise, let $f(x) = y_0$. Clearly f is discontinuous, since for every $r \in Y$ it takes on the value r on a dense set. However, we will show that for every nonempty $X \in \mathscr{F}$, $f(X) \in \mathscr{G}$. This will contradict the choice of $\mathscr{F}$ and $\mathscr{G}$ and will therefore complete the proof of Theorem 3. To prove our last assertion, suppose $X \in \mathscr{F}$ and X is nonempty. If X is a singleton then so is $f(X)$, so $f(X) \in \mathscr{G}$ by Claim 1. If not then by Claim 5, for every $r \in Y$, $X \cap X_{A_r} \neq \varnothing$. It follows that $f(X) = Y \in \mathscr{G}$. ■

For generalizations of Theorem 3, see **[1]**.

REFERENCES

1. Ciesielski, K., Dikranjan, D., and Watson, S., Functions characterized by images of sets, preprint.
2. Folland, G., *Real Analysis: Modern Techniques and their Applications*, New York: John Wiley and Sons, 1984.
3. Strichartz, R., *The Way of Analysis*, Boston: Jones and Bartlett Publishers, 1995.

Department of Mathematics and Computer Science
Amherst College
Amherst, MA 01002
djvelleman@amherst.edu

Probabilistic Pursuits on the Grid

A. M. Bruckstein, C. L. Mallows, and I. A. Wagner

1. INTRODUCTION: PROBABILISTIC PURSUIT. The paths of a sequence of a(ge)nts engaged in a sequence of continuous pursuits converge to the straight line between the origin and destination [**2**]. We consider a discrete setting where the a(ge)nts are only allowed to visit grid points and chase each other according to a probabilistic rule of motion, and prove a similar result: the average paths of ants in a chain of probabilistic pursuit converge rapidly to a straight line. This discrete model of pursuit leads to interesting results also in the context of linear and cyclic pursuits.

Assume that a sequence of ants $\mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2, \ldots$ are released from the origin at times $t = 0, \Delta, 2\Delta, \ldots,$ (Δ being an integer > 1), and each ant moves on the integer grid in the plane so that $\mathbf{A}_{n+1}$ chases or pursues $\mathbf{A}_n$ according to a probabilistic rule defined in the sequel. For sake of simplicity, consider that each ant measures time from its moment of release: if $\mathbf{A}_{n+1}$ is at time t of its motion (i.e., on the tth point of its trajectory), then $\mathbf{A}_n$ is at time $(t + \Delta)$. A pursuing ant $\mathbf{A}_{n+1}$ stays one unit of time at a grid point $A_{n+1}(t) = (x_{n+1}(t), y_{n+1}(t))$. Then it looks around, and decides where to move next according to the location $A_n(t + \Delta) = (x_n(t + \Delta), y_n(t + \Delta))$ of the pursued ant. Ant locations on the grid will be encoded as complex numbers: $A_n(t) \triangleq x_n(t) + jy_n(t)$, where $j = \sqrt{-1}$.

Probabilistic pursuit is defined by the following rule. $\mathbf{A}_{n+1}$ chooses its next position as one of its four nearest neighbor-points on the grid, under a probability distribution determined by its relative position with respect to the pursued ant. Thus

$$A_{n+1}(t+1) = A_{n+1}(t) + \delta_{n+1}(t+1), \tag{1}$$

where $\delta_{n+1}(\cdot)$ are random variables taking values in $\{1, -1, j, -j\}$ according to

$$\begin{aligned} \operatorname{Prob}\{\delta_{n+1}(t+1) = \operatorname{sign}(d_x)\} &= \frac{|d_x|}{d} \\ \operatorname{Prob}\{\delta_{n+1}(t+1) = j \cdot \operatorname{sign}(d_y)\} &= \frac{|d_y|}{d} \end{aligned} \tag{2}$$

where d_x, d_y are defined as

$$d_x = x_n(t+\Delta) - x_{n+1}(t)$$
$$d_y = y_n(t+\Delta) - y_{n+1}(t)$$

and $d = |dx| + |dy|$ is the "Manhattan distance" (the Manhattan norm of $x + jy$ is defined as $\|x + jy\| \overset{\text{def}}{=} \|x\| + \|y\|$) between successive ants (see Figure 1). If d drops to zero at some time during $\mathbf{A}_{n+1}$'s pursuit of $\mathbf{A}_n$, the ants merge and continue $\mathbf{A}_n$'s pursuit of $\mathbf{A}_{n-1}$. The preceding equations define a probabilistic pursuit in the complex plane, with pursuit steps biased according to the relative locations of the pursuer and pursued. The rule is trivial if $\Delta = 1$, since then the pursuing ant follows the leader exactly.

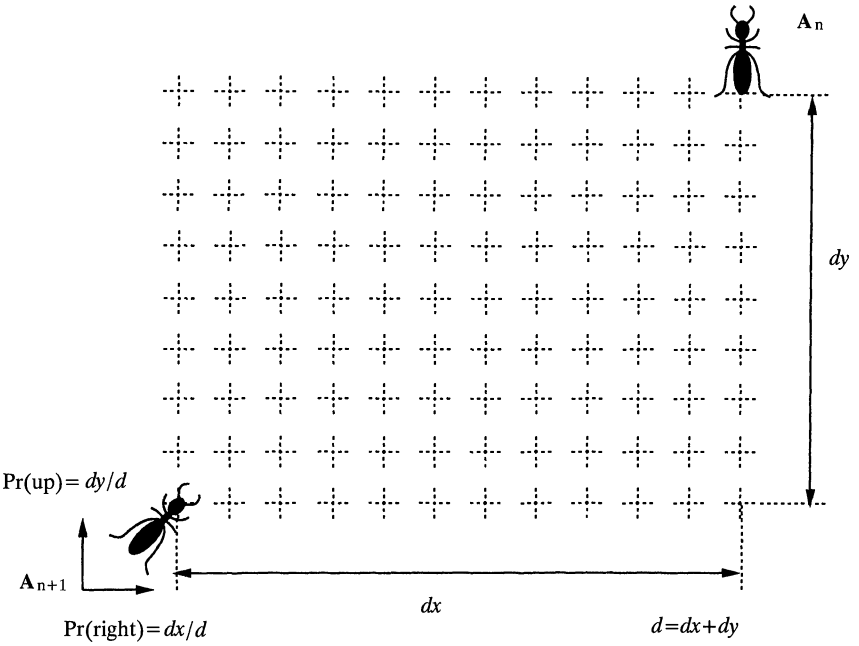
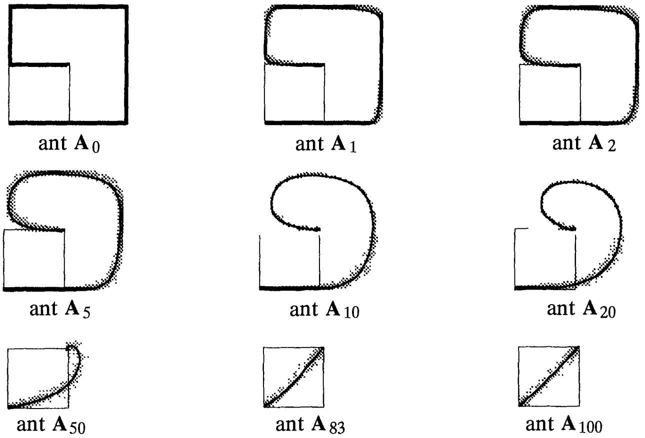


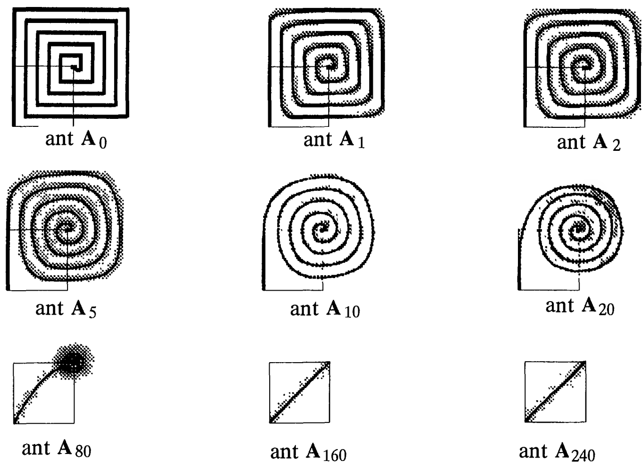
Figure 1. The probabilistic model for ant pursuits on $\mathbf{Z}^2$.

Figures 2–4 display simulation results of probabilistic pursuits for various initial trajectories. In each of these simulations we ran many pursuits with identical trajectories for $\mathbf{A}_0$, starting at $(0,0)$ and ending at some grid point (a, b). The figures show the distribution of locations visited by certain ants, the grey level of each pixel being proportional to the number of times the ant visited that location. The ensemble-averaged path of the sample ants is depicted as a bold curve.



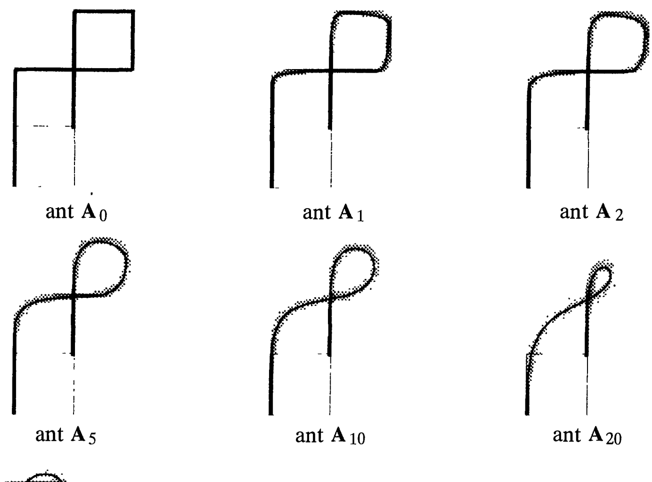
Probabilistic chain pursuit of 100 ants from (0, 0) to (20, 20)
Gray level - Distribution of sites visited by sample ants
Bold lines - the average path in 200 simulation runs
Initial Manhattan distance = 5

Figure 2. Probability distribution with a simple 'maze' initial path.



Probabilistic chain pursuit of 240 ants from (0, 0) to (20, 20)
Gray level - Distribution of sites visited by sample ants
Bold lines - the average path in 100 simulation runs
Initial Manhattan distance = 5

Figure 3. Probability distribution with yet another 'maze' initial path.



Probabilistic chain pursuit of 100 ants from (0, 0) to (20, 20)
Gray level - Distribution of sites visited by sample ants
Bold lines - the average path in 200 simulation runs
Initial Manhattan distance = 5

Figure 4. Probability distribution with a self-crossing initial path.

2. PATH CONVERGENCE TO STRAIGHT LINES. Assume that the first ant $\mathbf{A}_0$ travels along an arbitrary grid path from the origin to $a + jb$, where it stops (without loss of generality we assume that $a \geq 0$, $b \geq 0$). Then, for each $n \geq 0$, $\mathbf{A}_{n+1}$ pursues $\mathbf{A}_n$ following the probabilistic pursuit rule given by (1) and (2). Let

us define L_n as the (rectilinear) length of this path:

$$L_n = \sum_{t=0}^{T_n} \| A_n(t+1) - A_n(t) \|,$$

which equals T_n—the total number of steps in the path of the nth ant.

We shall show that the pursuit paths converge, in a sense, to the "straightest" line on the grid connecting the source 0 to the destination $a + jb$. This will be done in three stages: first we show that for any initial grid path taken by $\mathbf{A}_0$ the pursuit trajectories eventually become confined to the rectangle defined by 0 and $a + jb$, and are monotonic (of length $a + b$). Then we show that within the rectangle all monotonic paths have, in the limit, equal probability. This means that the points near the straight diagonal are more likely to be visited, and that the straight diagonal from 0 to $a + jb$ is the average path in the limit. Then we show that the average path converges to the straight line very fast.

2.1. The Pursuit Paths become Monotonic. We first show that the trajectory $A_n(t)$ eventually becomes monotonic. A discrete path is monotonic if it has no "backtracking"—that is, $\delta(t) \in \{1, j\}$ for all t during the pursuit.

Lemma 1. *L_n, the Manhattan path-lengths of ants engaged in probabilistic pursuit, is a positive, non-increasing (hence convergent) sequence.*

Proof: Since $T_n = L_n$, we show the claimed properties for T_n. Ant $\mathbf{A}_{n+1}$ starts its journey exactly Δ units of time after $\mathbf{A}_n$ has started. After T_n units of time, $\mathbf{A}_n$ stops at the destination and at this point $\mathbf{A}_{n+1}$ has made $T_n - \Delta$ steps along its trajectory. According to the probabilistic pursuit rules, the distance between ants can never increase, hence when $\mathbf{A}_n$ stops, its pursuer $\mathbf{A}_{n+1}$ is at a distance $\leq \Delta$ away from the destination. In the following $\Delta_f \leq \Delta$ units of time, $\mathbf{A}_{n+1}$ decreases its distance from the destination by exactly one per unit of time. Therefore we have

$$L_{n+1} = T_{n+1} = T_n - \Delta + \Delta_f \leq T_n - \Delta + \Delta = T_n = L_n$$

and since the sequence L_n is also bounded below by $a + b$, it converges. ■

We next claim that if the path-length of an ant is greater than $a + b$, there is a positive probability that the path-length of the next ant decreases.

Lemma 2

$$\text{Prob}\{L_{n+1} \leq L_n - 2 | L_n > a + b\} \geq \left(\frac{\Delta - 1}{\Delta}\right)^{L_0}$$

Proof: Since an ant starts at 0 and finally arrives at $a + jb$, it is clear that for all n we must have

$$\sum_{t=0}^{T_n} \delta_n(t) = a + jb.$$

From the definition of probabilistic pursuit we see that $\delta_n(t) \in \{\pm 1, \pm j\}$, and if $L_n > a + b$ (as we assume) the path of $\mathbf{A}_n$ is necessarily non-monotonic, that is: there exist times t_1, t_2 such that $\delta_n(t_2) = -\delta_n(t_1)$. Let us take (t_1, t_2) to be the earliest such interval, so that t_2 is the first time (after t_1) when $\mathbf{A}_n$ makes a "backtracking;" see Figure 5, in which we assume (without loss of generality) that at time t_1 the ant $\mathbf{A}_n$ moves to the left, then up, and at time t_2 to the right. Since

we require t_2 to be the first "backtracking," $\mathbf{A}_n$ moves upwards monotonically between $t_1 + 1$ and $t_2 - 1$, for h steps, where $h = t_2 - t_1 - 2$. Since $A_{n+1}(t)$ is Δ steps behind $\mathbf{A}_n$, at time t_1 it must be somewhere on the boundary of the square *RVQPTSWR*. Now from the figure it is clear under what conditions the distance between the ants decreases during the time interval (t_1, t_2). This happens if either (i) $A_{n+1}(t)$ is located to the left of *WT*, in which case the distance decreases at time t_1, or (ii) it is located to the right of V, in which case the distance decreases at time t_2. Also, if $A_{n+1}(t)$ is on *SPQ* the distance decreases sometime between t_1 and t_2. The only chance to preserve the distance is when $\mathbf{A}_{n+1}$ happens to be located on the arc *WRV* at time t_1; in this case $\mathbf{A}_{n+1}$ may first get to *PR*, and then follow $\mathbf{A}_n$ one step to the left of *PR* and later (after t_2) to the right, without ever shortening the distance between them. However this is not sure to happen. Wherever $\mathbf{A}_{n+1}$ starts from, there is the possibility that after it reaches *PR* it never makes a step to the left between t_1 and t_2. Let us denote by I the event "$\mathbf{A}_{n+1}$, once it has arrived on the line *PR*, stays there (at least) until time t_2". As explained previously

$$\operatorname{Prob}\{L_{n+1} < L_n\} \geq \operatorname{Prob}\{I\}.$$

To obtain a lower bound on the probability that I occurs, note that the probability that $\mathbf{A}_{n+1}$ does not move left in a certain time in (t_1, t_2), according to the

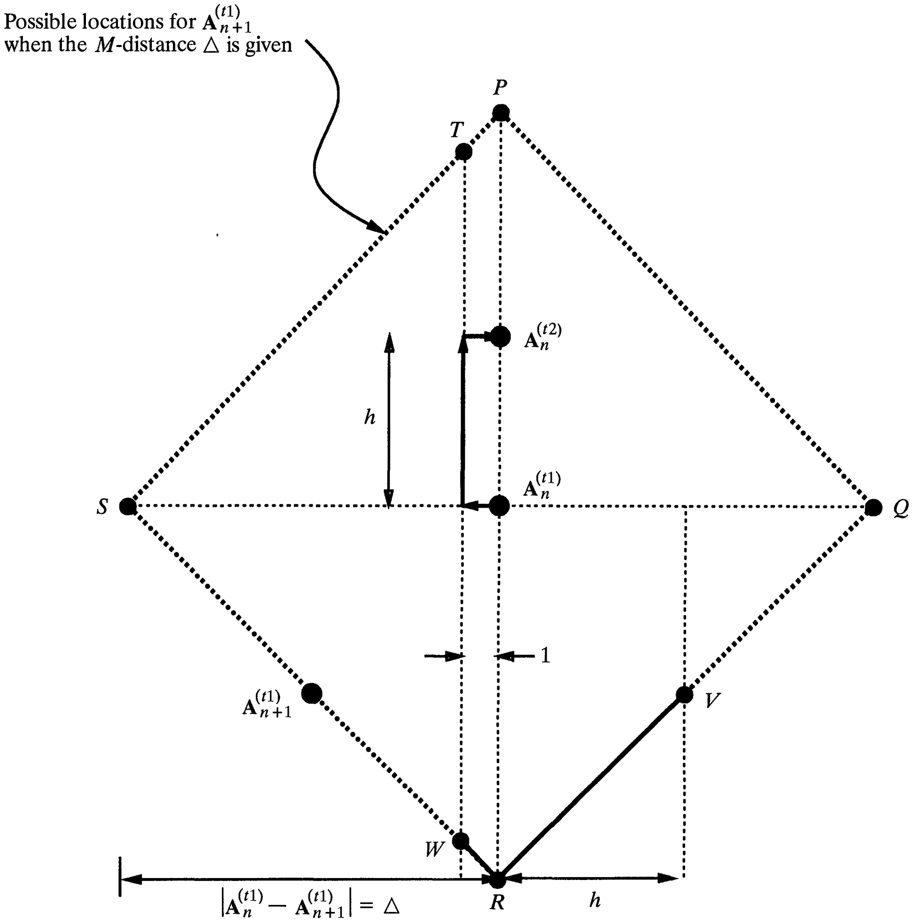


Figure 5. An illustration of a non-monotonic ant path.

probabilistic pursuit rule, is proportional to the ratio of d_y (($\Delta - 1$) in our case) to $d_x + d_y$ (Δ in our case). The event of staying on the line PR should repeat $t_2 - t_1$ times (or fewer if $\mathbf{A}_{n+1}$ arrives on the line PR later than t_1). Hence

$$\operatorname{Prob}\{I\} \geq \left(\frac{\Delta - 1}{\Delta}\right)^{t_2 - t_1}, \tag{3}$$

which is the probability that $\mathbf{A}_{n+1}$ stays on the line PR during an interval that is not longer than (t_1, t_2), given that $\mathbf{A}_n$ is hopping along the line TW. This effort by $\mathbf{A}_{n+1}$ is eventually rewarded at time t_2, when $\mathbf{A}_n$ turns right and the distance decreases by 2. Clearly,

$$t_2 - t_1 \leq T_n = L_n \leq L_0,$$

and hence the probability that the length of the $(n+1)$st path is shorter than that of the nth path by two (or more) units is bounded below by $((\Delta - 1)/\Delta)^{L_0}$. ∎

Note that if the distance between ants $\mathbf{A}_{n+1}$ and $\mathbf{A}_n$ drops, it drops in quanta of two if $\mathbf{A}_n$ is not stationary at $a + jb$. The proof of Lemma 2 also shows that chasing an ant that moves along a non-monotonic path induces a positive probability for a drop in the distance between the ants.

The next theorem shows that the pursuit path eventually becomes monotonic: L_n converges to $a + b$ with probability 1. In general, a sequence of random variables $\{X_n\}$ converges *with probability* 1 (or *almost surely*) to a value X (we write $X_n \xrightarrow{\text{a.s.}} X$) if, given $\epsilon, \delta > 0$, there exists an $n_0(\epsilon, \delta)$ such that for all $n > n_0$, $\operatorname{Prob}\{|X_n - X| < \delta\} > 1 - \epsilon$.

Theorem 1. *There exist constants $k_1, k_2 > 0$ such that, given $\epsilon > 0$, if*

$$n > n_0(\epsilon) = k_1 + k_2 \cdot \log\left(\frac{1}{\epsilon}\right)$$

then

$$\operatorname{Prob}\{L_n = a + b\} > 1 - \epsilon,$$

where L_n is the length of the path of $\mathbf{A}_n$ in a probabilistic pursuit from the origin to $a + jb$.

Proof: If $L_n > a + b$ then there must have been at most $s_0 = [L_0 - (a + b)]/2 - 1$ ants in the sequence $\mathbf{A}_0, \ldots, \mathbf{A}_n$ for which a drop (of 2) in the distance to the pursued ant occurred, since a decrease in the distance between consecutive ants implies a decrease in the path length of the pursuing ant. Hence, there were at least $n - s_0$ ants with no decrease in distance. Lemma 2 ensures that each ant path can be viewed as the outcome of an experiment in which the distance-drop event occurs with a probability of at least $p = ((\Delta - 1)/\Delta)^{L_0}$. A sequence of ants engaged in a probabilistic pursuit is a series of trials, with outcomes that are either a "success"—a drop in the inter-ant distance (which has a probability at least p), or a "failure"—the distance does not change. Define A to be the event "*s_0 or fewer distance-drops in a chain of n ants*".

$$\begin{aligned}\operatorname{Prob}\{L_n > a + b\} = \operatorname{Prob}\{A\} &= \sum_{s=0}^{s_0} \operatorname{Prob}\{s \text{ successes up to } n\} \\ &\leq (1-p)^n + \binom{n}{1}(1-p)^{n-1} + \cdots + \binom{n}{s_0}(1-p)^{n-s_0}\end{aligned}$$

$$= (1-p)^n \sum_{s=0}^{s_0} \binom{n}{s} (1-p)^{-s}$$

$$\leq (1-p)^n \binom{n}{s_0} \sum_{s=0}^{s_0} (1-p)^{-s} \quad (\text{for } n > 2s_0)$$

$$\leq (1-p)^n \binom{n}{s_0} \frac{s_0}{(1-p)^{s_0}} \leq (1-p)^n n^{s_0} C_1 = C_1 q^n n^{C_2}.$$

Here C_1, C_2, and $q < 1$ are constants, independent of n and ϵ. Since $\lim_{n\to\infty} C_1 \cdot q^n \cdot n^{C_2} = 0$, there exist constants C_3, C_4 such that

$$\text{for all } n > C_3, \quad C_1 \cdot q^n \cdot n^{C_2} < C_4 \cdot q^{n/2}$$

and in order to get

$$\text{Prob}\{A\} < C_4 \cdot q^{n/2} < \epsilon$$

it is sufficient to have

$$n > \frac{2 \log C_4}{\log \frac{1}{q}} + \frac{2}{\log \frac{1}{q}} \log\left(\frac{1}{\epsilon}\right). \qquad \blacksquare$$

2.2. The Stationary Path-Distribution is Uniform. The paths followed by successive ants form a Markov chain, with the state-space being all paths from the origin to $a + jb$. Theorem 1 ensures that all paths longer than $m = a + b$ are transitory. If we restrict to paths of length exactly m, we shall show that the chain is irreducible and aperiodic (and therefore ergodic), with the stationary distribution being uniform. If the initial path is monotone, the rule (2) has the following interpretation, which greatly simplifies some of the proofs we offer:

> Suppose we have a supply of black and white balls, and a series of urns $\mathbf{U}_0, \mathbf{U}_1, \mathbf{U}_2, \ldots$, which initially are all empty. At time $t = 1, 2, \ldots, a + b$ an agent $\mathbf{A}_0$ places a ball, either white or black, into $\mathbf{U}_0$. At each time $\Delta, \Delta + 1, \ldots$, agent $\mathbf{A}_1$ takes a ball at random from $\mathbf{U}_0$ (which at time Δ contains Δ balls) and places it in $\mathbf{U}_1$. At each time $2\Delta, 2\Delta + 1, \ldots$, agent $\mathbf{A}_2$ takes a ball at random from $\mathbf{U}_1$ and places it in $\mathbf{U}_2$, and so on. For each urn, the number of balls it contains starts by rising from zero to Δ, stays there a while, and then decreases to zero.

This description is equivalent to that of probabilistic pursuit, if we take a white ball for a right-step and a black ball for an up-step, and identify the position $\mathbf{A}_n(t)$ with $w + jv$ where w (respectively, v) is the total number of white (respectively, black) balls this agent has seen by time t. The number of white (black) balls in urn $\mathbf{U}_{n-1}$ corresponds to the x (y) position of $\mathbf{A}_{n-1}$ relative to $\mathbf{A}_n$. If $\mathbf{A}_n(t) = w + jv$ and $\mathbf{A}_{n-1}(t) = w + jv + x + jy$, so that the urn $\mathbf{U}_{n-1}$ contains x white and y black balls, then the probability that $\mathbf{A}_n$ chooses a white ball (so that $\mathbf{A}_n(t+1) = w + 1 + jv$) is just $x/(x+y)$.

Let S be the set of monotonic paths from the origin to $a + jb$, and let $\mathscr{M}$ be the Markov chain with state-space S and transition probabilities induced by the probabilistic pursuit procedure.

We first show that $\mathscr{M}$ is irreducible.

Lemma 3. *For any two paths $s, s' \in S$ there is a sequence of positive-probability transitions that leads from s to s'.*

Proof: One can interpret a monotonic path from 0 to $a + jb$ as a sequence of $a + b$ characters from the set $\{u, r\}$, where r refers to a "right" move and u to an "up" move. There are exactly a r's and b u's. It is easy to see that if, in the target's

path s, there is a u at time t, followed by an r at time $t+1$, then there is a positive probability that the pursuer's path s' will be equal to s with the only exception that s' has an r at time t and a u at time $t+1$. The set S of monotonic paths is closed under such "flip" operations—given a path $s \in S$, any other path in S can be reached from s by a sequence of (positive probability) "flip" transitions. Hence the chain is irreducible. ∎

It is easy to see that $\mathscr{M}$ is aperiodic:

Lemma 4. *For any path $s \in S$, $p_{ss} > 0$.*

Proof: There is always some positive probability that the pursuer follows the pursued's path exactly. ∎

Now we show

Lemma 5. *The uniform distribution over S is stationary.*

Proof: The number of different paths from the origin to $a + jb$ is

$$|S| = \binom{m}{a}.$$

For the uniform distribution of paths, the position at time t (starting from the origin at $t = 0$) is $x + jy$ (where $x + y = t$) with probability

$$\operatorname{Prob}\{x|m,t,a\} = \frac{1}{|S|}\binom{t}{x}\binom{m-t}{a-x}. \tag{4}$$

This is the hypergeometric distribution, which governs the number of white balls (x) in a random sample of t balls chosen from an urn that contains a white and b black balls. Thus we can generate a random path by choosing balls sequentially at random from an urn that initially has a white and b black balls.

Next consider the case when $t + \Delta < a + b$. Suppose the path of the pursued ("target") ant, $\mathbf{A}_1$, is chosen uniformly from S, e.g., by drawing from an urn with a white and b black balls, and moving right on white and up on black. Using the "urn" representation, we can obtain the distribution over all possible paths for the kth ant by considering a sequence of urns $\mathbf{U}_0, \mathbf{U}_1, \ldots, \mathbf{U}_k, \ldots$ with the black and white balls being moved downstream according to the following rule:

> Start with $\mathbf{U}_0$ containing a white and b black balls. At each time unit draw a ball at random from $\mathbf{U}_0$ and place it into $\mathbf{U}_1$ until Δ balls are accumulated there. Then also start moving randomly chosen balls from $\mathbf{U}_1$ to $\mathbf{U}_2$ until Δ balls are in $\mathbf{U}_2$ and so forth.

The distribution of paths for the kth ant is given by the distribution of ball-color sequences seen entering the urn $\mathbf{U}_k$ in this process. Disregarding the color of balls, by symmetry all $(a+b)!$ sequences of balls are equally probable to appear as inputs to $\mathbf{U}_k$. Hence the

$$\frac{(a+b)!}{a!b!} = \binom{a+b}{a}$$

possible sequences of black and white balls are also equiprobably seen entering the kth urn. ∎

The property we have just proved is strongly related to the concept of exchangeability, defined as follows (see [**6**, pp. 97–105]): A countable sequence of events

$V_1, V_2, \dots$ is *exchangeable* if for any possible choice $1 \le i_1 < i_2 < \cdots < i_k$ of k subscripts, $\text{Prob}(V_{i_1} \cap V_{i_2} \cap \cdots \cap V_{i_k}) = p_k$ depends only on k but not on the actual subscripts i_j. If the event V_i is defined as "a white ball enters the last urn at time i", then the probability of having a such events does not depend on the order in which they occur, hence the sequence is exchangeable and all paths are equiprobable.

The preceding result is quite general. In fact, if we take a sequence of urns with a white and b black balls in the first one and move them downstream, choosing balls at random from $\mathbf{U}_i$ to be placed into $\mathbf{U}_{i+1}$, according to any given schedule ensuring that all balls pass through each urn, then all the possible color sequences of balls entering each urn have the same probability. This shows that for monotone pursuits one can vary the inter-ant intervals arbitrarily, and the paths of the ants engaged in pursuit will be uniformly distributed if the first ant chooses a path at random from $(0,0)$ to (a,b). This also generalizes to higher dimensions (= more colors for balls). Thus the paths generated by this rule are also governed by a uniform stationary distribution.

From Lemmas 3, 4, and 5 we have

Theorem 2. *$\mathscr{M}$ is an ergodic Markov chain and its unique stationary distribution is uniform.*

Two immediate corollaries of Theorem 2 are:

Corollary 1. *Assuming stationarity, the average path is the straight line from* 0 *to* $a + jb$.

Proof: A standard result for the hypergeometric distribution (4) is that $\mathbf{E}[x|m,t,a] = ta/m$. ■

Corollary 2. *Assuming stationarity, ants are usually very near the average path.*

Proof: For the hypergeometric distribution (4), the variance of x is

$$\mathbf{V}[x|m,t,a] = t(m-t)ab/(m-1)m^2.$$

Thus if $a = \alpha m$, $b = \beta m$, and $t = \tau m$ (where $\alpha + \beta = 1$) we have:

$$\mathbf{V}[x(t)] = m\alpha\beta\tau(1-\tau) + O(1).$$

Suppose m is large. We can bound the probability that at time t the ant is outside a region of width m^ϵ around the average, ϵ being a number in $(\frac{1}{2}, 1)$. Using Chebyshev's inequality,[1]

$$\begin{aligned}
&\text{Prob}\left\{\left|x(t) - \frac{at}{m}\right| \ge m^\epsilon\right\} \\
&\quad = \text{Prob}\left\{\left|x(t) - \frac{at}{m}\right|^2 \ge m^{2\epsilon}\right\} \\
&\quad \le \frac{\mathbf{V}[x(t)]}{m^{2\epsilon}} = \alpha\beta\tau(1-\tau)m^{1-2\epsilon} + O(m^{-2\epsilon}) \xrightarrow{\text{as } m \to \infty} 0.
\end{aligned}$$

[1]Chebyshev's inequality ([5, p. 376]) says: let X be a random variable with expected value $\mathbf{E}[X]$ and variance $\mathbf{V}[X]$. Then $\text{Prob}\{(X - \mathbf{E}[X])^2 \ge \alpha\} \le \mathbf{V}[X]/\alpha$ for any $\alpha > 0$.

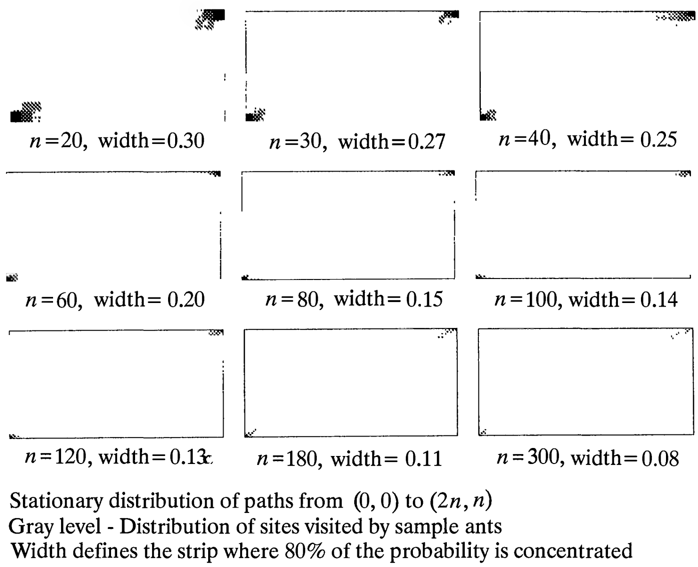


Figure 6. Line widths for stationary distribution when $a = 2b$.

The normalized width of the strip with positive probability is $n^{\epsilon}/\alpha m$, which clearly converges to zero when $m \to \infty$. See Figure 6 for the line width in the stationary distributions for various values of m. ■

3. CONVERGENCE TO THE STRAIGHT LINE IS FAST. We now show that the average of the ant-paths converges to the straight line between source and destination exponentially fast.

In the following, we ignore the initial non-monotonic transient, and assume that the leading ant $\mathbf{A}_0$ executes an arbitrary monotonic path. Let us define a new entity $\mathbf{D}_n$ (a determin-ant?) which progresses along the average path of $\mathbf{A}_n$, i.e. such that at each time t, $\mathbf{D}_n(t) = \mathbf{E}[\mathbf{A}_n(t)]$. Then

$$D_{n+1}(t+1) = D_{n+1}(t) + \frac{D_n(t+\Delta) - D_{n+1}(t)}{\Delta}. \tag{5}$$

To justify this equation, note that the expectation of the step made by $\mathbf{A}_{n+1}$ at time t is

$$\mathbf{E}[A_{n+1}(t+1)] - [A_{n+1}(t)] = \frac{\mathbf{E}[A_n(t+\Delta)] - \mathbf{E}[A_{n+1}(t)]}{\Delta}.$$

Let us denote the average path of the ant $\mathbf{A}_n$ by the complex vector $\mathbf{d} = (d(0), d(1), d(2), \dots, d(m))$, where $m = a + b$, and denote the path of the pursuing ant by $\mathbf{d}' = (d'(0), d'(1), d'(2), \dots, d'(m))$. We measure the distance between these two paths by the maximum distance between any of their components, i.e.,

$$\operatorname{dist}(\mathbf{d}, \mathbf{d}') = \max_{0 \le i \le m} |d(i) - d'(i)|,$$

where $|\cdot|$ stands for the Euclidean distance. Now we can show that the average path approaches its linear limit exponentially fast.

Theorem 3

$$\operatorname{dist}(\mathbf{d}_n, \mathbf{d}_\infty) \le \frac{m(m-1)}{\alpha^{m-\Delta}}(1 - \alpha^{m-\Delta})^n, \tag{6}$$

where $\alpha = (\Delta - 1)/\Delta$.

Proof: First we show that the limit average path, $\mathbf{d}_\infty$, is indeed the straight line. We can write the evolution equations as

$$\begin{aligned} 0 \le t \le m - \Delta: \qquad & d'(t+1) - d'(t) = \frac{d(t+\Delta) - d'(t)}{\Delta} \\ m - \Delta < t < m: \qquad & d'(t+1) - d'(t) = \frac{d(m) - d'(t)}{m - t} \end{aligned} \tag{7}$$

with boundary conditions

$$d(0) = d'(0) = 0, \qquad d(m) = d'(m) = a + jb,$$

where the denominators represent the Manhattan distances between $\mathbf{A}_n$ and $\mathbf{A}_{n+1}$. This distance is initially Δ, and stays constant until $\mathbf{A}_n$ reaches $a + jb$, whereupon the distance decreases by one per unit of time. Hence we can relate the vectors $\mathbf{d}$ and $\mathbf{d}'$ in the following way:

$$\begin{aligned} d'(0) &= d(0) \\ \Delta d'(1) + (1-\Delta)d'(0) &= d(\Delta) \\ \Delta d'(2) + (1-\Delta)d'(1) &= d(\Delta+1) \\ &\vdots \\ \Delta d'(m-\Delta+1) + (1-\Delta)d'(m-\Delta) &= d(m) \\ \hline (\Delta-1)d'(m-\Delta+2) + (2-\Delta)d'(m-\Delta+1) &= d(m) \\ (\Delta-2)d'(m-\Delta+3) + (3-\Delta)d'(m-\Delta+2) &= d(m) \\ &\vdots \\ 2d'(m-1) + (-1)d'(m-2) &= d(m) \\ d'(m) &= d(m). \end{aligned} \tag{8}$$

A fixed point of this linear iterative process is a vector $\mathbf{d}$ such that $\mathbf{d}' = \mathbf{d}$. In such a vector, $d(t+1) - d(t)$ must be constant for all t. Otherwise, assume that there is a solution for which the sequence $d(t+1) - d(t)$ is not constant, and denote $x(t) = \Re d(t)$; the same argument holds for $y(t) = \Im d(t)$. Denote by t_0 the smallest integer in $[0, m-2]$ such that the difference $x(t_0+1) - x(t_0)$ is an extremum—either a minimum or a maximum. This difference is necessarily nonnegative since the path is monotonic. From (7) it follows that

$$\begin{aligned} x(t_0+1) - x(t_0) &= \frac{x(t_0+\delta) - x(t_0)}{\delta} = \frac{1}{\delta}\sum_{k=1}^{\delta}(x(t_0+k) - x(t_0+k-1)) \\ &= \frac{1}{\delta}\sum_{k=1}^{\delta}|x(t_0+k) - x(t_0+k-1)|. \end{aligned}$$

Hence

$$\min_{1\le k\le\delta}|x(t_0+k)-x(t_0+k-1)|<|x(t_0+1)-x(t_0)|$$
$$<\max_{1\le k\le\delta}|x(t_0+k)-x(t_0+k-1)|,$$

where $\delta=\min\{\Delta, m-t_0\}>1$. The last inequality is strict since not all the differences are equal. But this contradicts our assumption that $x(t_0+1)-x(t_0)$ is an extremum. Moreover, t_0 cannot equal $m-1$, since then both the minimum and maximum would occur at the same index, contradicting the assumption that the sequence is non-constant.

Since $d(0)$ and $d(m)$ are not affected by the iterative process, the vector $\mathbf{d}_n$ converges to a limit that is a sequence of points equi-spaced on the straight line from $d(0)$ to $d(m)$.

We next show that the distance from the limit decreases exponentially fast. The set of difference equations (8) can be written as:

$$\mathbf{\Phi d}'=\mathbf{\Psi d},$$

where the matrices $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are

$$\mathbf{\Phi}_{(m+1)\times(m+1)} \qquad \qquad \longleftarrow \quad \Delta \quad \longrightarrow$$

$$=\begin{array}{cc}
\left(\begin{array}{cccc|ccccc}
1 & 0 & & \cdots & & & & \cdots & 0\\
1-\Delta & \Delta & 0 & \cdots & & & & \cdots & 0\\
0 & 1-\Delta & \Delta & 0 & \cdots & & & \cdots & 0\\
\vdots & & & \ddots & & & & & \vdots\\
\cdot & \cdot & \cdot & 1-\Delta & \Delta & \cdot & & \cdot & \vdots\\
\hline
\cdot & \cdot & \cdot & \cdot & 2-\Delta & \Delta-1 & & \cdot & \vdots\\
\vdots & & & & & & & & \vdots\\
\vdots & & & & & & -1 & 2 & \vdots\\
0 & & \cdots & & & & & 0 & 1
\end{array}\right)
&
\begin{array}{c}
\\ \\ \\ \\ \\ \uparrow \\ \Delta-1 \\ \downarrow \\ \\
\end{array}
\end{array}$$

and

$$\mathbf{\Psi}_{(m+1)\times(m+1)}=
\begin{array}{cc}
\begin{array}{c} \longleftarrow \; \Delta \; \longrightarrow \qquad\qquad\qquad\qquad \end{array} & \\
\left(\begin{array}{cccc|cccccc}
1 & 0 & \cdots & 0 & 0 & 0 & & 0 & \cdots & 0\\
0 & 0 & \cdots & 0 & 1 & 0 & & 0 & \cdots & 0\\
0 & 0 & \cdots & 0 & 0 & 1 & & 0 & \cdots & 0\\
\vdots & & & & & & \ddots & & & \vdots\\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & 1 & 0\\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & 0 & 1\\
\hline
\vdots & & & & & & & & & 1\\
\vdots & & & & & & & & & \vdots\\
0 & & \cdots & & & & & & & 1
\end{array}\right)
&
\begin{array}{c}
\\ \\ \\ \\ \\ \\ \downarrow \\ \Delta-1 \\ \downarrow
\end{array}
\end{array}$$

Note that $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are independent of the specific path. Hence, the dynamics of the averaged ant-paths is described by

$$\mathbf{d}'=\mathbf{\Phi}^{-1}\cdot\mathbf{\Psi}\cdot\mathbf{d}=\mathbf{P}\cdot\mathbf{d}$$

i.e., a fixed matrix operator repeatedly acting on the average ant-path vector. Let us now sketch the form of this operator and derive a bound on its second-largest eigenvalue.

With some algebraic manipulations, it can be found that

$\mathbf{\Phi}^{-1} =$

$$\left(\begin{array}{ccccc|ccc}
1 & & & & & & & 0 \\
\alpha & \beta & & & & & & \\
\alpha^2 & \alpha\beta & \beta & & & & & \\
\alpha^3 & \alpha^2\beta & \alpha\beta\ \beta & & & & & \\
\vdots & & & \ddots & & & & \\
\alpha^{m-\Delta+1} & \alpha^{m-\Delta}\beta & \cdots & \beta & & & & \\
\hline
\alpha^{m-\Delta+1}\left(\frac{\Delta-2}{\Delta-1}\right) & \alpha^{m-\Delta}\beta\left(\frac{\Delta-2}{\Delta-1}\right) & \cdots & \beta\left(\frac{\Delta-2}{\Delta-1}\right) & \frac{1}{\Delta-1} & & & \\
\alpha^{m-\Delta+1}\left(\frac{\Delta-3}{\Delta-1}\right) & \alpha^{m-\Delta}\beta\left(\frac{\Delta-3}{\Delta-1}\right) & \cdots & \beta\left(\frac{\Delta-3}{\Delta-1}\right) & \left(\frac{\Delta-3}{(\Delta-1)(\Delta-2)}\right) & \frac{1}{\Delta-2} & & \\
\vdots & & & & & & \ddots & \\
\alpha^{m-\Delta+1}\left(\frac{1}{\Delta-1}\right) & \alpha^{m-\Delta}\beta\left(\frac{1}{\Delta-1}\right) & \cdots & \beta\left(\frac{1}{\Delta-1}\right) & \left(\frac{1}{(\Delta-1)(\Delta-2)}\right) & \cdots & & \frac{1}{2} \\
0 & 0 & \cdots & & & & & 0 \quad 1
\end{array}\right)$$

with $\alpha = (\Delta - 1)/\Delta$ and $\beta = 1/\Delta$, and hence

$$\longleftarrow \qquad \Delta \qquad \longrightarrow$$

$$\mathbf{P} = \mathbf{\Phi}^{-1}\mathbf{\Psi} = \left(\begin{array}{cc|ccccc}
1 & 0\cdots 0 & 0 & & & & 0 \\
\alpha & 0\cdots 0 & \beta & & & & \\
\alpha^2 & 0\cdots 0 & \alpha\beta & \beta & & & \\
\vdots & 0\cdots 0 & \alpha^2\beta & \alpha\beta & \beta & & \\
\alpha^{m-\Delta} & 0\cdots 0 & \vdots & & & \ddots & 0 \\
\hline
\alpha^{m-\Delta+1} & 0\cdots 0 & \alpha^{m-\Delta}\beta & \cdots & & & \beta \\
\alpha^{m-\Delta+1}\left(\frac{\Delta-2}{\Delta-1}\right) & 0\cdots 0 & \alpha^{m-\Delta}\beta\left(\frac{\Delta-2}{\Delta-1}\right) & \cdots & & & \frac{2}{\Delta} \\
\alpha^{m-\Delta+1}\left(\frac{\Delta-3}{\Delta-1}\right) & 0\cdots 0 & \alpha^{m-\Delta}\beta\left(\frac{\Delta-3}{\Delta-1}\right) & \cdots & & & \frac{3}{\Delta} \\
\vdots & 0\cdots 0 & \vdots & & & & \vdots \\
\alpha^{m-\Delta+1}\left(\frac{1}{\Delta-1}\right) & 0\cdots 0 & \alpha^{m-\Delta}\beta\left(\frac{1}{\Delta-1}\right) & \cdots & & & \frac{\Delta-1}{\Delta} \\
0 & 0\cdots 0 & 0 & \cdots & & 0 & 1
\end{array}\right)$$

Note that the row sums of $\mathbf{P}$ are all 1.

Since the fixed point of the process $\mathbf{d}' = \mathbf{P} \cdot \mathbf{d}$ is the straight line from $d(0)$ to $d(m)$, and is independent of the entries $d(1), d(2), \ldots, d(m-1)$ in the initial $\mathbf{d}$, we know that as n tends to infinity, $\mathbf{P}^n$ approaches the form of two non-zero columns on left and right, all other entries being zeroes. In order to analyze the rate of convergence of this process, let us bound the value of $p_{ij}^{(n)}$, the (i, j)th entry in $\mathbf{P}^n$. An observation we need for this purpose is that the sum of the central $m - 1$ entries in any row of $\mathbf{P}$ is bounded from above:

$$\sum_{k=1}^{m-1} p_{ik} \leq 1 - \alpha^{m-\Delta},$$

with equality achieved at the $(m - \Delta + 1)$th row of $\mathbf{P}$. Using this observation and the fact that the top and bottom entries in the $m - 1$ central columns of $\mathbf{P}^n$ are

zero for all n, we have the following recursive argument:

$$\begin{aligned}
p_{ij}^{(n)} &= \sum_{k=1}^{m-1} p_{ik} \cdot p_{kj}^{(n-1)} \\
&\le \left(\sum_{k=1}^{m-1} p_{ik} \right) \cdot \max_{0<k<m} \left\{ p_{kj}^{(n-1)} \right\} \\
&\le (1 - \alpha^{m-\Delta}) \cdot \max_{0<k<m} \left\{ p_{kj}^{(n-1)} \right\} \\
&\le (1 - \alpha^{m-\Delta})^2 \cdot \max_{0<k<m} \left\{ p_{kj}^{(n-2)} \right\} \\
&\;\vdots \\
&\le (1 - \alpha^{m-\Delta})^n .
\end{aligned} \tag{9}$$

Hence, the magnitudes of all the entries of $\mathbf{P}^n$ except for those in the leftmost and rightmost columns tend to zero rather quickly. Now let us consider the 0th and mth columns. Due to the special structure of $\mathbf{P}$ and the inequalities (9) we have that for all i, $0 \le i \le m$,

$$\begin{aligned}
p_{i0}^{(n)} &= p_{i0}^{(n-1)} + \sum_{k=1}^{m-1} p_{ik}^{(n-1)} \cdot p_{k0} \\
&\le p_{i0}^{(n-1)} + (m-1)(1 - \alpha^{m-\Delta})^{n-1},
\end{aligned}$$

and

$$\begin{aligned}
\left| p_{i0}^{(n)} - p_{i0}^{(\infty)} \right| &\le (m-1) \sum_{k=n}^{\infty} (1 - \alpha^{m-\Delta})^k \\
&= \frac{m-1}{\alpha^{m-\Delta}} (1 - \alpha^{m-\Delta})^n,
\end{aligned}$$

i.e., the leftmost entries of $\mathbf{P}^n$ approach their limit values exponentially fast, too. A similar argument holds for the entries of the rightmost column. We conclude that the effect of the initial conditions (i.e., of $d(1), d(2), \ldots, d(m-1)$ in $\mathbf{d}_0$) decays exponentially fast, and the average ant path converges to the straight line as expressed by (6). ■

4. RELATED TOPICS. We now consider several extensions to the probabilistic pursuit model.

4.1. Probabilistic Linear Pursuit. Consider two ants, the first of which, $\mathbf{A}_0$, is happily hopping along a straight line parallel to the y-axis: $A_0(t) = r + jt$, where r is a constant. A second ant, $\mathbf{A}_1$, is chasing $\mathbf{A}_0$, and both are traveling at the same speed. Using our probabilistic pursuit model, one can get an equation for the average trajectory of $A_1(t)$, similar to the corresponding deterministic results found in [**1**, pp. 251–253] and [**4**, pp. 113–127].

Theorem 4. *If $\mathbf{A}_0$ is launched from $(r,0)$ at time 0 and is going upwards at speed 1, and if $\mathbf{A}_1$ is launched from $(0,0)$ at time 0 and is pursuing $\mathbf{A}_0$ according to the probabilistic pursuit model, the average behavior of $A_1(t)$ is described by the curve*

$$y(x) = \frac{\log\left(\dfrac{r-x}{r}\right)}{\log\left(\dfrac{r-1}{r}\right)} - x.$$

Proof: Since the behavior of the ants can be described by the equations

$$\begin{aligned} A_0(t) &= r + jt \\ A_1(0) &= 0 \\ A_1(t) &= A_1(t-1) + \delta(t), \end{aligned} \tag{10}$$

where $\delta(t)$ is the random variable defined in (2). Since the rectilinear distance between them is always r, the average y-coordinate of $\mathbf{A}_1$ at time t is

$$y_t = y_{t-1} + \frac{t - 1 - y_{t-1}}{r}$$

with initial condition $y_0 = 0$. Substituting $\alpha = (1 - \frac{1}{r})$, $\beta = \frac{1}{r}$, and using the fact that $x_0 = y_0 = 0$, it turns out that

$$\begin{aligned} y_t &= \alpha y_{t-1} + \beta(t-1) = \alpha(\alpha y_{t-2} + \beta(t-2)) + \beta(t-1) = \cdots \\ &= \beta \sum_{k=1}^{t-1} \alpha^{k-1}(t-k) = \beta\alpha^{t-1} \sum_{k-1}^{t-1} k\alpha^{-k} = r\left(1 - \frac{1}{r}\right)^t + t - r. \end{aligned}$$

Solving (10) for x_t, we get

$$\begin{aligned} x_t &= \alpha x_{t-1} + 1 = \alpha^2 x_{t-2} + \alpha + 1 = \cdots \\ &= \alpha^t x_0 + \sum_{k=0}^{t-1} \alpha^k = \frac{1 - \alpha^t}{1 - \alpha} = r - r\left(1 - \frac{1}{r}\right)^t, \end{aligned}$$

hence

$$y(x) = \frac{\log\left(\frac{r - x}{r}\right)}{\log\left(\frac{r-1}{r}\right)} - x. \qquad \blacksquare$$

This result is quite similar to the one obtained for continuous linear pursuit [**1**, p. 251]:

$$y(x) = \frac{(x-r)^2}{4c} - \frac{c}{2}\log(r - x) + c',$$

where c, c' are constants. The difference is explained by the different measures of distance involved: in our model the ant moves toward its target with a constant speed, maintaining a constant Manhattan distance to it, but the length of the average step it takes in the direction of the target varies, while in [**1**] the pursuit is carried out with constant Euclidean velocity pointed at the chased ant. Note that the Euclidean ant is asymptotically at distance $r/2$ behind its target, while the Manhattan ant never decreases its distance below r. See Figure 7 for a graphic comparison of pursuit path induced by these two models.

4.2. Probabilistic Cyclic Pursuit. Assume that $\mathbf{A} = \{\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_n\}$ is a set of ants, chasing each other cyclically, that is: $\mathbf{A}_1$ is chasing $\mathbf{A}_0$, $\mathbf{A}_2$ is chasing $\mathbf{A}_1$, etc., and $\mathbf{A}_0$ is chasing $\mathbf{A}_n$. The set $\mathbf{A}$ begins at positions $A(0)$ at time $t = 0$ and then evolves on according to the probabilistic pursuit rules defined in the previous section.

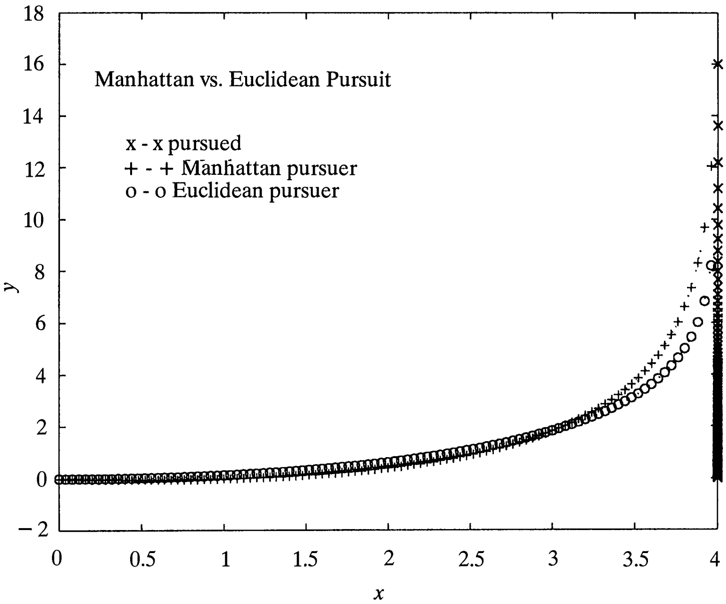


Figure 7. Comparison of the Manhattan and Euclidean models of pursuit.

Denote by C_t the *Manhattan circumference* of the set A:

$$C_t = \sum_{i=0}^{n} \|A_{i+1}(t) - A_i(t)\|$$

where $\|u - v\|$ denotes the Manhattan distance between points u and v. In **[2]** and **[3]** it was shown that ants engaged in deterministic cyclic pursuit always converge to a point of mutual encounter (and all captures are almost always simultaneous, see **[7]**). Here we shall show that the ants reach a *limit cycle*, each ant being not more than one unit of distance away from its chaser.

Theorem 5. *Ants engaged in cyclic probabilistic pursuit with initial distances* $d_1, d_2, \ldots, d_n$ *converge to a limit cycle with circumference* $C_\infty = \sum_{i=0}^{n} (d_i \bmod 2)$. *Moreover, this convergence is exponentially fast: for any given* $\epsilon > 0$, *if* $t > t_0(\epsilon) = O(\log(\frac{1}{\epsilon}))$ *then* $\text{Prob}\{C_t = C_\infty\} > 1 - \epsilon$.

Proof: Inter-ant distances never increase in probabilistic pursuit, hence C_t is a non-increasing positive, hence convergent, sequence. Arguments similar to those in the proof of Lemma 2 show that whenever the distance between two ants is greater than 1 there is a positive probability, bounded from below, for a decrease (by 2) in this distance, provided the pursued ants' path is non-monotonic. But, in the case of cyclic pursuit, the paths of all ants are obviously non-monotonic, since they all have infinite length and are confined to the "bounding box" of the initial configuration. Hence C_∞ must correspond to a limiting pursuit configuration in which all distances are less than 2, proving the first part of the assertion of the theorem.

To prove that the convergence is exponentially fast, note that, as in the proof of Lemma 2, the inter-ant distance drops by 2 with probability higher than

$$\left(\frac{1}{2}\right)^{\text{length of non-monotonic run}} > \left(\frac{1}{2}\right)^{C_0}$$

(since C_0 is an obvious upper bound on all such runs) each time a non-monotonic run occurs in the pursued ant's trajectory. But this happens at least once every C_0 steps (since the ant must stay within a bounding box of Manhattan perimeter of at most C_0). Hence we have

$$\operatorname{Prob}\{C_{t+C_0} \leq C_t - 2 | C_t > C_\infty\} \geq \left(\frac{1}{2}\right)^{C_0}.$$

In order to get $\operatorname{Prob}\{C_t = C_\infty\} > 1 - \epsilon$, we must (as in Theorem 1) have t of the order of $\log(1/\epsilon)$. ■

The limit cycle may be a polygon with (up to) $n + 1$ vertices, as long as the length of each edge is exactly one unit; see Figure 8 for an example. Such a polygon is stable since in this case each ant $\mathbf{A}_{i+1}$ "replaces" the pursued one $\mathbf{A}_i$, the overall shape is preserved. Figures 9–14 exhibit simulation examples of the probabilistic cyclic pursuit. For each of the initial configurations we show the evolution of the probability distribution calculated over a large number of experiments, as well as the actual ant locations in a single experiment. It would be interesting to investigate the relation between the shape of the initial polygon whose vertices are $A_i(0)$, $i = 0, 1, \ldots, n$, and the shape of the limit cycle.

5. CONCLUDING REMARKS. Many of the results of this paper continue to hold when the lag Δ is not held constant, but is allowed to vary from one ant to the next. We could also allow for the chasing ant to be guided by an ant other than the one immediately ahead. To achieve the asymptotic results, we need only ensure that eventually the current ant is many generations removed from the first one. Also we need to have $\Delta \geq 2$ infinitely often at each stage of the walk.

The results discussed in this paper can be generalized to three (or more) dimensional space. The probability of $\mathbf{A}_{n+1}$ moving along each axis will, in this

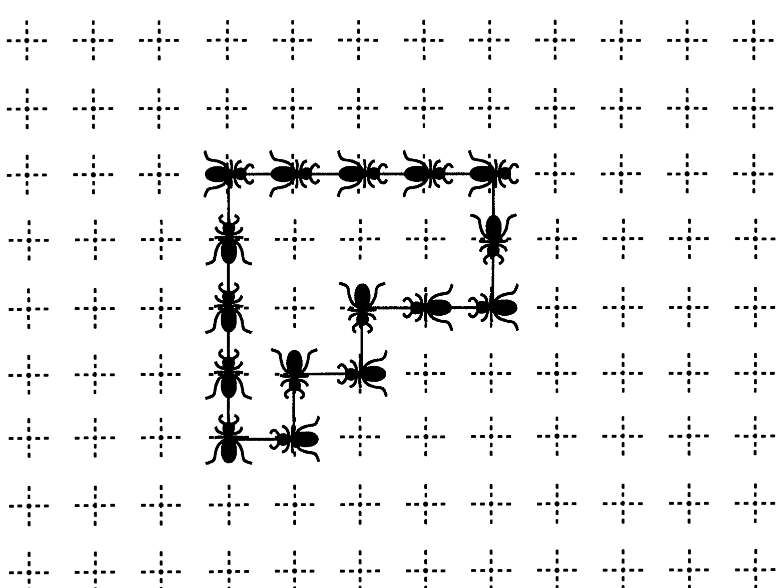


Figure 8. A possible limit cycle for a cyclic pursuit.

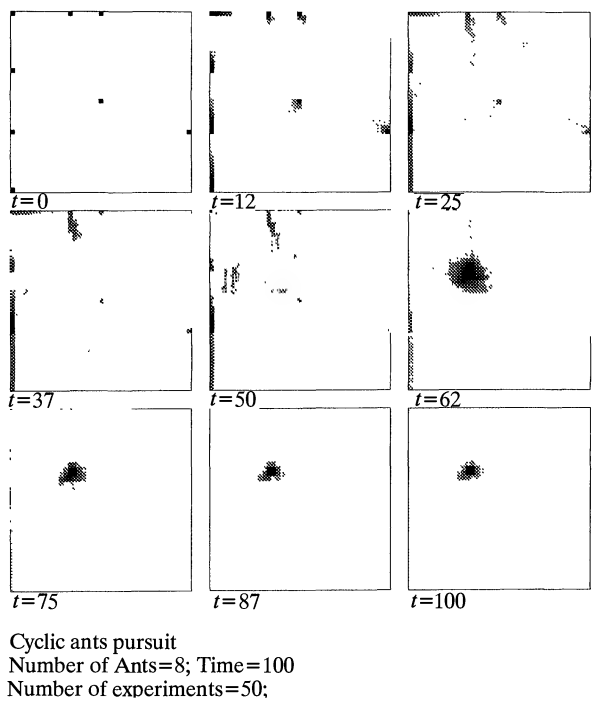


Figure 9. Probability distribution in cyclic pursuit–initial configuration 1.

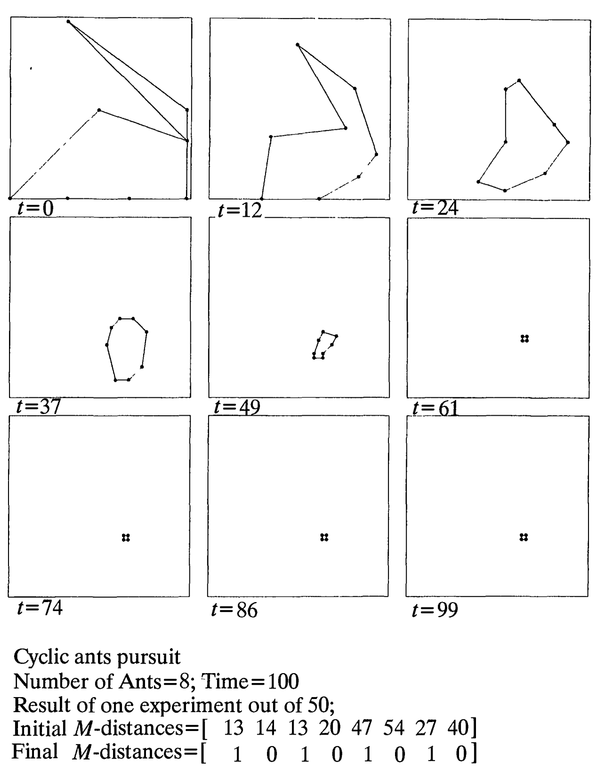
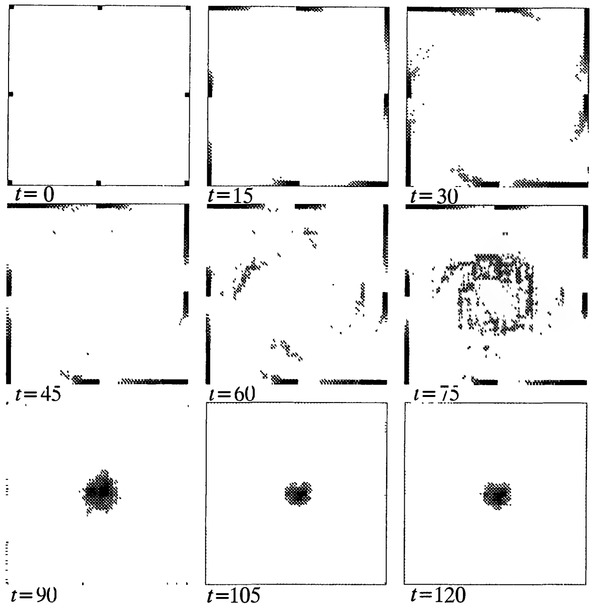
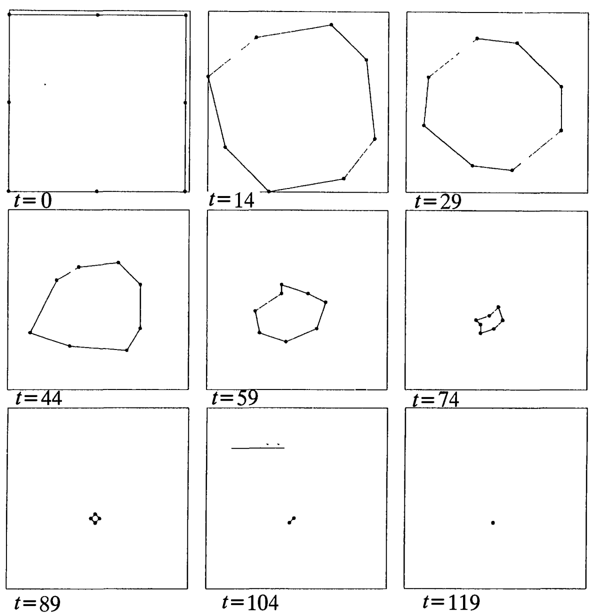


Figure 10. A single run of cyclic pursuit–initial configuration 1.



Cyclic ants pursuit
Number of Ants=8; Time=120
Number of experiments=50;

Figure 11. Probability distribution in cyclic pursuit–initial configuration 2.



Cyclic ants pursuit
Nymber of Ants=8; Time=120
Result of one experiment out of 50;
Initial M-distances=[20 20 20 20 20 20 20 20]
Final M-distances=[0 0 0 0 0 0 0 0]

Figure 12. A single run of cyclic pursuit–initial configuration 2.

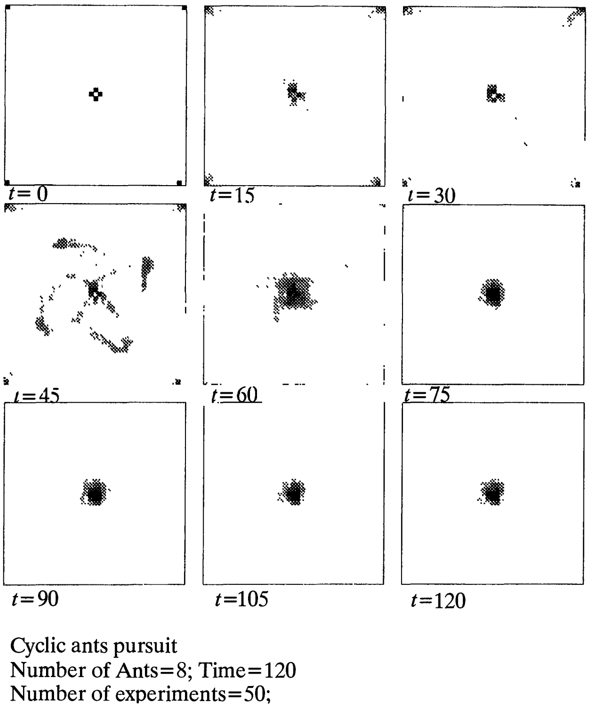


Figure 13. Probability distribution in cyclic pursuit–initial configuration 3.

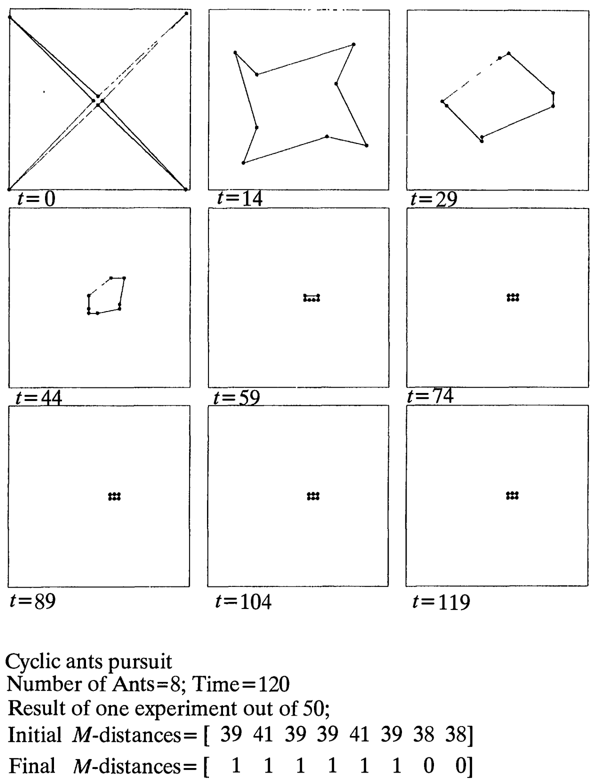


Figure 14. A single run of cyclic pursuit–initial configuration 3.

case, be proportional to the projection of the vector $A_n - A_{n+1}$ along this axis.

Ants obeying the probabilistic pursuit model have the property of moving, on the average, in the same direction as a continuous pursuit. However, their speed is not constant since it depends on the location of the chaser relative to the target. To overcome this problem, for purposes of approximating continuous pursuit, one might consider the following Euclidean probabilistic rule of pursuit:

$$\begin{aligned}
P_x &= \operatorname{Prob}\{\delta_{n+1}(t+1) = \operatorname{sign}(d_x)\} &&= \frac{1}{2}\cdot\frac{|d_x|}{\sqrt{dx^2+dy^2}} \\
P_y &= \operatorname{Prob}\{\delta_{n+1}(t+1) = j\cdot\operatorname{sign}(d_y)\} &&= \frac{1}{2}\cdot\frac{|d_y|}{\sqrt{dx^2+dy^2}} \\
P_0 &= \operatorname{Prob}\{\delta_{n+1}(t+1) = 0\} &&= 1-\frac{1}{2}\cdot\frac{|d_x|+|d_y|}{\sqrt{d_x^2+d_y^2}}
\end{aligned} \tag{11}$$

where $d_x = x_n(t+\Delta) - x_{n+1}(t)$ and $d_y = y_n(t+\Delta) - y_{n+1}(t)$ are defined as before. The additional "Euclidization" factor does not affect the average direction of the chaser, but does normalize its velocity to $\frac{1}{2}$, independent of the target's location: it is easy to verify that $P_x + P_y + P_0 = 1$ and that $(P_x^2 + P_y^2)^{1/2} = \frac{1}{2}$. It is an open question whether some or all of our results hold for this model. The main difficulty is caused by the non-zero probability for the chaser to stay at its current location, which means that the pursuit distance is not monotonically decreasing, as it is in the Manhattan case.

ACKNOWLEDGMENT. We wish to thank Bob Holt of AT & T-Bell Labs for his help in "debugging" an early version of this paper, Amir Dembo of the Technion for his contribution to the simplification of the proof of Lemma 5, and the anonymous referees for their enlightening comments.

REFERENCES

1. G. Boole, *A Treatise on Differential Equations*, Chelsea Publishing Company, London, 1859.
2. A. M. Bruckstein, Why the Ant Trails Look So Straight and Nice, *The Mathematical Intelligencer*, vol. 15, No. 2, pp. 59–62, 1993.
3. A. M. Bruckstein, N. Cohen, and A. Efrat, *Ants, Crickets and Frogs in Cyclic Pursuit*, Center for Intelligent Systems Report #9105, Technion, Israel, 1991.
4. H. T. Davis, *Introduction to Nonlinear Differential and Integral Equations*, Dover, NY, 1962.
5. R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics*, Addison-Wesley, Reading, MA, 1991.
6. N. L. Johnson and S. Kotz, *Urn Models and Their Applications—An Approach to Modern Discrete Probability Theory*, John Wiley and Sons, NY, 1977.
7. T. J. Richardson, *Cyclic Pursuit: Captures and Limits*, AT & T Bell Laboratories Technical Memo, December 1991.

Alfred M. Bruckstein
Department of Computer Science
Technion City, Haifa 32000, Israel
freddy@cs.technion.ac.il

Colin L. Mallows
Statistics Research Department
AT & T Bell Labs, Murray Hill, NJ 07974
clm@research.att.com

Israel A. Wagner (corresponding author)
IBM–Haifa Research Laboratory
MATAM, Haifa 31905, Israel
wagner@haifasc3.vnet.ibm.com
http://www.cs.technion.ac.il/~wagner

Hipparchus, Plutarch, Schröder, and Hough

Richard P. Stanley

1. HIPPARCHUS AND PLUTARCH. Plutarch was a Greek biographer and philosopher from Chaeronea, who was born before A.D. 50 and died after A.D. 120. He is best known for his *Parallel Lives*, which inspired such Renaissance writers as Montaigne, Shakespeare, Dryden, and Rousseau. His many other works have been gathered together under the name *Moralia*, "a collection of comparatively short treatises and dialogues which cover an immense range of subjects, literary, ethical, political, and scientific" [**21**, p. 8]. Part of the *Moralia* consists of the *Table-Talk*, "a collection of dialogues purporting to reproduce the after-dinner conversation of Plutarch and his friends and relatives on various occasions" [**20**, p. 2]. In the *Table-Talk* [**20**, VIII.9, 732] appears the following statement:

> Chrysippus says that the number of compound propositions that can be made from only ten simple propositions exceeds a million. (Hipparchus, to be sure, refuted this by showing that on the affirmative side there are 103,049 compound statements, and on the negative side 310,952.)

Chrysippus (*c*. 280–207 B.C.) came to Athens around 260 and became a leading Stoic philosopher. Hipparchus was a Greek astronomer (*c*. 190–after 127 B.C.) from Nicaea in Bithynia (now Iznik, Turkey) who spent much of his life at Rhodes. He was perhaps the greatest astronomer of antiquity. He is most famous for his discovery of the precession of the equinoxes, based on his own observations and those of Timocharis 160 years earlier. For further information on the work of Hipparchus, see [**19**, Book I, E], [**32**]. Hipparchus was an excellent mathematician (though for a contrary view see [**33**, p. 211]); he was the first person to make systematic use of trigonometry, and he was probably the inventor of stereographic projection. However, for many centuries no one was able to make sense of the statement of Plutarch. For instance, T. L. Heath [**12**, vol. 2, p. 256], a standard older authority on Greek mathematics, says of Plutarch's statement that "it seems impossible to make anything of these figures," while the more recent authority O. Neugebauer [**19**, p. 338] states that Plutarch's statement "[has], however, so far eluded a satisfactory explanation." Similarly W. and M. Kneale [**16**, p. 162], authorities on the history of logic, remark that "It is difficult to make any satisfactory sense of the passage." N. L. Biggs [**2**, p. 113] notes the paucity of combinatorial computations by the ancient Greeks and referring to Plutarch's passage says that "the most interesting of them is also the most mysterious." A number of eminent mathematicians and historians of mathematics, such as M. Cantor, J. Tropfke, S. Günther, and E. Artin, have attempted to understand Plutarch's statement without success. An attempt to reconstruct Hipparchus' procedure appears in [**1**], though it will be apparent from our discussion that this attempt is incorrect. Another incorrect speculation appears in [**30**, p. 63].

2. SCHRÖDER. Friedrich Wilhelm Karl Ernst Schröder was a German logician who was born in Mannheim on November 25, 1841, and died in Karlsruhe on June

16, 1902. He passed the doctoral exam at the University of Heidelberg in 1862 and had positions in Zurich (at the Eidgenössische Polytechnikum), Karlsruhe, Pforzheim, and Baden-Baden, before accepting a post as full professor at Karlsruhe in 1876. Schröder worked mainly on the foundations of mathematics, notably with combinatorics, the theory of functions of a real variable, and mathematical logic. He was one of the first persons to accept Cantor's ideas in set theory and was one of the developers of mathematical logic in the second half of the nineteenth century. Schröder is best known to combinatorialists for his paper **[25]**, in which he discusses four "bracketing problems." The first two problems concern the bracketing or parenthesization of a string of letters that we may assume to be all identical, say the letter x. The second two problems are analogues of the first two where the string of letters is replaced by a set of elements. We will discuss only the first two problems here.

The formal definition of a bracketing is the following. First, x itself is considered to be a bracketing. Recursively define a bracketing to be a sequence $B = (B_1, \dots, B_k)$, where $k \geq 2$ and each B_i is a bracketing. We represent the bracketing B as a parenthesized string of x's. Thus, think of B as a k-ary product $(B_1)(B_2) \cdots (B_k)$. If some B_i is the single letter x, then we remove the parentheses surrounding B_i for clarity of notation. Thus, for example, the bracketing

$$(xx)((xxxxx)x(xx))(xx(xx)) \tag{1}$$

represents a way of multiplying 14 x's whose last operation was a ternary operation $(B_1)(B_2)(B_3)$, where $B_1 = xx$, $B_2 = (xxxxx)x(xx)$, and $B_3 = xx(xx)$, and similarly for B_1, B_2, and B_3. There are exactly eleven bracketings of four letters, namely,

$$xxxx \quad (xx)xx \quad x(xx)x \quad xx(xx) \quad (xxx)x \quad x(xxx)$$
$$((xx)x)x \quad (x(xx))x \quad (xx)(xx) \quad x((xx)x) \quad x(x(xx)).$$

Note that the last five of these are built up entirely from *binary* operations and are therefore called *binary bracketings*.

There are three fundamental equivalent ways to represent a bracketing in addition to a parenthesized string discussed above: as *plane trees*, *polygon dissections*, and *Łukasiewicz words*. We now briefly describe these alternative representations. If B is a bracketing, then we first define the plane tree $\tau(B)$ corresponding to B. If B consists of a single letter, then $\tau(B)$ is a single root vertex. If $B = (B_1, \dots, B_k)$ then $\tau(B)$ consists of a root vertex (drawn at the top), with subtrees $\tau(B_1), \dots, \tau(B_k)$, drawn in that order from left to right. Thus, the key property defining a plane tree is that the subtrees of every vertex are linearly ordered. For instance, the plane tree corresponding to the bracketing of equation (1) is shown in Figure 1. Note that a binary bracketing corresponds to a *binary plane tree*, i.e., a plane tree for which every non-endpoint vertex has exactly two successors.

Next we consider polygon dissections. Let P be a convex polygon. A *dissection* of P is obtained by drawing some diagonals that don't intersect in their interiors. Thus, P is divided up into regions that are themselves convex polygons. In particular, if P has m sides and we draw $m - 3$ such diagonals (the maximum number possible), then we obtain a dissection for which every region is a triangle; such dissections are called *triangulations*. We now explain how to associate a plane tree $\tau(D)$ with a polygon dissection D. We associate with the "degenerate" polygon with just two vertices a single root vertex. Now fix once and for all an edge e of the polygon P, called the *root edge*. In a given dissection D, the edge e is contained in a unique polygon Q that is a region of D. Let $k + 1$ be the number of

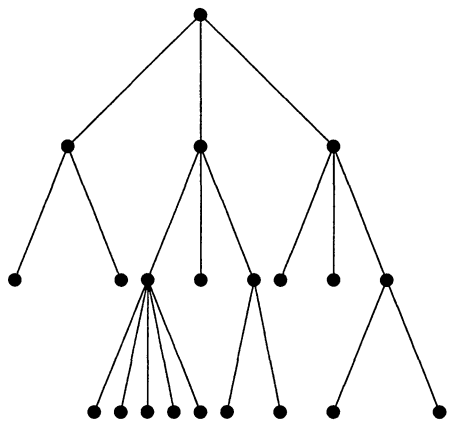


Figure 1. A plane tree.

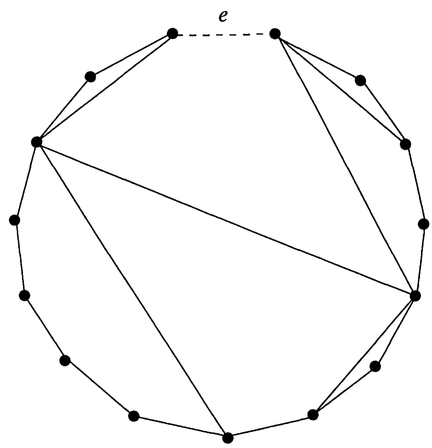


Figure 2. A polygon dissection.

edges of Q. If we remove the edge e and the interior of Q from D, then we are left with dissections $D_1, D_2, \ldots, D_k$ of k polygons (some possibly with just two vertices), reading counterclockwise from e along the boundary of Q, such that D_i and D_{i+1} intersect at a single vertex for $1 \leq i \leq k-1$. Define recursively $\tau(D)$ to be the plane tree whose subtrees of the root are $\tau(D_1), \ldots, \tau(D_k)$ in that order. Note that if P has $n+1$ vertices, then $\tau(D)$ has n endpoints. Figure 2 shows the polygon dissection corresponding to the tree of Figure 1.

Finally we consider Łukasiewicz words. The letters of such words come from the alphabet $A = \{x_0, x_1, x_2, \ldots\}$. The *weight* $\delta(x_i)$ of a letter x_i is defined by $\delta(x_i) = i - 1$. A word $y_1 y_2 \cdots y_m$ made of letters from A is said to be a *Łukasiewicz word* if $\delta(y_1) + \cdots + \delta(y_j) \geq 0$ for $1 \leq j \leq m-1$, and $\delta(y_1) + \cdots + \delta(y_m) = -1$. Thus, $y_m = x_0$. The set of all Łukasiewicz words is called the *Łukasiewicz language* [**17**, Ch. 11.3]. To obtain a Łukasiewicz word $\omega(\tau)$ from a plane tree τ, do a depth-first (preorder) search through the tree. By definition, this is a linear ordering $\delta(\tau) = v_1, v_2, \ldots, v_p$ of the vertex set of τ defined recursively by $\delta(\tau) = v, \delta(\tau_1), \ldots, \delta(\tau_k)$, where v is the root of τ, and $\tau_1, \ldots, \tau_k$ are the

subtrees of v (in that order). Define

$$\omega(\tau) = x_{\deg(v_1)} x_{\deg(v_2)} \cdots x_{\deg(v_k)},$$

where $\deg(v_i)$ denotes the degree (number of successors or children) of vertex v_i. For instance, the Łukasiewicz word corresponding to the plane tree of Figure 1 is

$$x_3 x_2 x_0^2 x_3 x_5 x_0^6 x_2 x_0^2 x_3 x_0^2 x_2 x_0^2.$$

Note that since our bracketings B do not allow unary operations, the plane tree $\tau(B)$ has no vertices of degree one, and the corresponding Łukasiewicz word does not involve the letter x_1.

The correspondences we have established are easily seen to yield the following result.

Proposition. (a) *Let $s(n)$ denote the total number of bracketings of a string of n letters. Then $s(n)$ is also equal to* (*i*) *the number of plane trees with no vertex of degree one and with n endpoints*, (*ii*) *the number of dissections of a convex* $(n+1)$*-gon, and* (*iii*) *the number of Łukasiewicz words with no x_1's and with n x_0's.*

(b) *Let $b(n)$ denote the number of binary bracketings of a string of n letters. Then $b(n)$ is also equal to* (*i*) *the number of binary plane trees with n endpoints* (*and hence with $2n-1$ vertices*), (*ii*) *the number of triangulations of a convex* $(n+1)$*-gon, and* (*iii*) *the number of Łukasiewicz words with n x_0's and $n-1$ x_2's* (*and with no other letters*); *such words, usually with the last x_0 deleted, are sometimes called* Dyck words.

We are now ready to explain the contribution of Schröder to these bracketing problems. Schröder's first problem asks for the number $b(n)$ of binary bracketings of a string of n letters. Using a generating function argument, Schröder derives the formula (stated slightly differently)

$$b(n) = \frac{1}{n}\binom{2n-2}{n-1}.$$

Thus $b(n)$ is just the *Catalan number* C_{n-1}, for which an enormous literature exists. For some further information and references, see **[11]**, **[14]**. A list of about fifty combinatorial interpretations of Catalan numbers will appear in [**31**, Exercise 6.17] and is available on the World Wide Web at
http://www-math.mit.edu/~rstan/ec/ec.html.

Schröder's second problem asks for the total number $s(n)$ of bracketings of a string of n letters. Schröder's main result on his second problem is the generating function

$$\sum_{n\geq 1} s(n)x^n = \frac{1}{4}\left(1 + x - \sqrt{1-6x+x^2}\right). \tag{2}$$

He also gives the values (with the typographical error 145 for $s(5) = 45$)

$$(s(1), \ldots, s(10)) = (1,1,3,11,45,197,903,4279,20793,103049). \tag{3}$$

Perhaps the quickest way to obtain equation (2) is the following. Let y denote the left-hand side. The recursive definition of bracketing is equivalent to the formula

$$y = x + y^2 + y^3 + y^4 + \cdots = x + \frac{y^2}{1-y}. \tag{4}$$

Multiplying by $1 - y$ yields the quadratic equation

$$2y^2 - (1+x)y + x = 0. \tag{5}$$

One of the solutions is spurious, and the other one is just the right-hand side of (2).

The numbers $s(n)$ are now called *Schröder numbers*. Schröder does not mention any other combinatorial interpretations of Schröder numbers, nor does he give a single outside reference. Let us point out some additional references. The problem of counting the triangulations of a convex polygon was raised by Segner [**26**] and solved (anonymously) by Euler [**9**]. The connection between bracketings and plane trees was known to Cayley [**4**]. The bijection between plane trees and polygon dissections appears in Etherington [**8**], with a sequel by Erdélyi and Etherington in [**7**]. The bijection between bracketings and Łukasiewicz works is essentially the "reverse Polish notation" or "parenthesis-free notation" developed by the Polish logician Jan Łukasiewicz (1878–1956). He came upon the idea of this notation in 1924 and first published it in 1929, as explained in [**18**, p. 180, footnote 3]. The connection between reverse Polish notation and enumerative combinatorics appears in a pioneering paper of George Raney [**22**].

There is now a considerable literature on Schröder numbers and related numbers. To get into this literature, see [**3**], [**15**, p. 55], [**23**], [**27**], and [**34**]. Let us also mention that it is easy to obtain a simple recurrence relation [**5**], [**6**, p. 57] for the Schröder numbers that allows them to be computed rapidly. Namely, differentiate (5) with respect to x and solve for y' to obtain

$$y' = \frac{y-1}{4y-1-x} = \frac{(x-3)y - x + 1}{x^2 - 6x + 1},$$

the latter equality a consequence of the quadratic equation (5). Hence

$$(x^2 - 6x + 1)y' - (x-3)y + x - 1 = 0.$$

Expanding the left-hand side in a power series in x and setting the coefficient of x^n equal to 0 yields

$$(n+2)s(n+2) - 3(2n+1)s(n+1) + (n-1)s(n) = 0, \quad n \geq 1. \tag{6}$$

No direct combinatorial proof of this formula was known until D. Foata and D. Zeilberger, after reading an earlier version of this paper, found such a proof [**10**].

3. HOUGH. The stage is now set for the *dénouement*. The astute reader may have already anticipated it by comparing Plutarch's cryptic statement with the values (3) of the Schröder numbers. In January 1994 David Hough (1949–), a graduate student at George Washington University (who decided only in 1992 that he would pursue a career in mathematics), noticed that the mysterious number 103,049 of Plutarch, i.e., the number of compound propositions that can be formed from ten simple propositions, is just the tenth Schröder number! Hough learned about Plutarch's statement from [**30**, Exercise 1.45]. Hough's discovery strongly suggests that Hipparchus was carrying out a calculation equivalent to the modern calculation of the number of bracketings of a string of ten letters. However, it remains to determine exactly what Hipparchus and Plutarch meant by a "compound proposition." In Stoic logic, compound propositions are built up from simple ones using such connectives as "and," "or," and "if . . . then" [**16**, Ch. III.5]. This does not seem like enough information to pinpoint precisely what Hipparchus had in mind.

We can also ask how Hipparchus computed the number 103,049. As noted in [**24**, p. 101], this number is much too large to have been computed by a direct enumeration of all the cases. Moreover, it is highly unlikely that Hipparchus was aware of the sophisticated recurrence (6). More probable is that Hipparchus used the "obvious" recurrence (equivalent to equation (4))

$$s(n) = \sum_{i_1 + \cdots + i_k = n} s(i_1) \cdots s(i_k), \quad n \geq 2, \tag{7}$$

where the sum ranges over all ways to write n as an (ordered) sum of $k \geq 2$ positive integers. The sum on the right-hand side of equation (7) in the case $n = 10$ has 511 terms. There are only 41 "essentially different" terms, corresponding to the 41 partitions of 10 into a least two parts, i.e., the 41 ways to write 10 as an *unordered* sum of at least two positive integers. If the terms of the sum are grouped according to the partition of 10 to which they correspond, it is still necessary to count the number of ways of ordering each partition. For instance, the partition $3 + 2 + 2 + 1 + 1 + 1$ has 60 orderings of its terms, thus contributing the amount $60s(3)s(2)^2s(1)^3$ to the sum (7). We cannot but admire Hipparchus' ability to compute the Schröder number $s(10)$ at a distant time when not even a remotely similar accurate computation is known. For further information about combinatorics in ancient times, see [**2**], [**24**].

The number 310,952 in Plutarch's statement, i.e., the number of compound propositions that can be formed from ten simple propositions "on the negative side," remains an enigma. Many possible variants of plane trees have been looked at without success. Moreover, Neil Sloane has verified that the numbers 310,952 and 103,049 + 310,952 = 414,001 do not appear anywhere in the valuable tables [**28**]. Thus the mystery of Plutarch's statement remains at most half solved.

ACKNOWLEDGMENT. The research was partially supported by NSF grant DMS-9500714. I am grateful to Judith Grabiner, Wilbur Knorr, and four anonymous referees for providing invaluable suggestions and references.

REFERENCES

1. K.-R. Biermann and J. Mau, Überprüfung einer frühen Anwendung der Kombinatorik in der Logik, *J. Symbolic Logic* **23** (1958), 129–132.
2. N. L. Biggs, The roots of combinatorics, *Historia Mathematica* **6** (1979), 109–136.
3. J. Bonin, L. W. Shapiro, and R. Simion, Some q-analogues of the Schröder numbers arising from combinatorial statistics on lattice paths, *J. Stat. Planning and Inference* **34** (1993), 35–55.
4. A. Cayley, On the analytical form called trees, Part II, *Philos. Mag.* (4) **18** (1859), 374–378.
5. L. Comtet, Calcul pratique des coefficients de Taylor d'une fonction algébrique, *Enseignement Math.* **10** (1964), 267–270.
6. L. Comtet, *Advanced Combinatorics*, Reidel, Dordrecht/Boston, 1974.
7. A. Erdélyi and I. M. H. Etherington, Some problems of non-associative combinations (2), *Edinburgh Math. Notes* **32** (1940), 7–12.
8. I. M. H. Etherington, Some problems of non-associative combinations (1), *Edinburgh Math. Notes* **32** (1940), 1–6.
9. L. Euler, Summarium, *Novi Commentarii academiae scientiarum Petropolitanae* **7** (1758/59), 13–15. Reprinted in *Opera Omnia* (1) **26** (1953), xvi–xviii.
10. D. Foata and D. Zeilberger, A classic proof of a recurrence for a very classical sequence, preprint. Available from the website http://www.math.temple.edu/~zeilberg/mamarim/mamarimhtml/classic.html.
11. M. Gardner, Catalan numbers, Mathematical Games, *Scientific American* **234** (June, 1976), pp. 120–125, and bibliography on p. 132. Reprinted (with an Addendum) in Chapter 20 of *Time Travel and Other Mathematical Bewilderments*, W. H. Freeman, New York, 1988.

12. T. L. Heath, *A History of Greek Mathematics*, Oxford University Press, Oxford; 1921; reprinted by Dover, New York, 1981.
13. T. L. Heath, *A Manual of Greek Mathematics*, Oxford University Press, Oxford, 1931; reprinted by Dover, New York, 1963.
14. P. Hilton and J. Pedersen, Catalan numbers, their generalizations, and their uses, *Math. Intelligencer* **13** (Spring, 1991), 64–75.
15. M. Klazar, On *abab*-free and *abba*-free set partitions, *Europ. J. Combinatorics* **17** (1996), 53–68.
16. W. Kneale and M. Kneale, *The Development of Logic*, Oxford University Press, Oxford, 1962, 1971.
17. M. Lothaire, *Combinatorics on Words*, Encyclopedia of Mathematics and Its Applications, vol. 17, Addison-Wesley, Reading, Massachusetts, 1983.
18. J. Łukasiewicz, *Selected Works* (L. Borkowski, ed.), North-Holland, Amsterdam, 1970.
19. O. Neugebauer, *A History of Ancient Mathematical Astronomy*, vol. 1, Springer-Verlag, Berlin/Heidelberg/New York, 1975.
20. Plutarch, *Moralia*, vol. IX (introduction by E. L. Minar, Jr.), Loeb Classical Library, Harvard University Press, Cambridge, Massachusetts, 1961.
21. Plutarch, *The Rise and Fall of Athens: Nine Greek Lives*, translated by I. Scott-Kilvert, Penguin, London, 1960.
22. G. N. Raney, Functional composition patterns and power series reversion, *Trans. Amer. Math. Soc.* **94** (1960), 441–451.
23. D. G. Rogers and L. W. Shapiro, Deques, trees and lattice paths, in *Combinatorial Mathematics VIII* (K. L. McAvaney, ed.), Lecture Notes in Mathematics, no. 884, Springer-Verlag, Berlin, pp. 293–303.
24. A. Rome, Procédés anciens de calcul des combinaisons, *Ann. Soc. sci. Bruxelles*, ser. A **50** (1930), 97–104.
25. E. Schröder, Vier combinatorische Probleme, *Z. für Math. Physik* **15** (1870), 361–376.
26. A. de Segner, Enumeratio modorum, quibus figurae planae rectilineae per diagonales dividuntur in triangula, *Novi Commentarii academiae scientiarum Petropolitanae* **7** (1758/59), 203–209.
27. L. W. Shapiro and A. B. Stephens, Bootstrap percolation, the Schröder numbers, and the n-kings problem, *SIAM J. Discrete Math.* **4** (1991), 275–280.
28. N. J. A. Sloane and S. Plouffe, *The Encyclopedia of Integer Sequences*, Academic Press, San Diego, 1995.
29. R. Stanley, Differentiably finite power series, *European J. Combinatorics* **1** (1980), 175–188.
30. R. Stanley, *Enumerative Combinatorics*, vol. 1, Wadsworth & Brooks/Cole, Belmont, CA, 1986; second printing, Cambridge University Press, 1996.
31. R. Stanley, *Enumerative Combinatorics*, vol. 2, Cambridge University Press, Cambridge, in preparation.
32. G. J. Toomer, Hipparchus, in *Dictionary of Scientific Biography* (C. C. Gillipsie, editor in chief), vol. XV, supplement I, Scribner's, New York, 1978, pp. 207–224.
33. B. L. van der Waerden, *Geometry and Algebra in Ancient Civilizations*, Springer-Verlag, Berlin, 1983.
34. J. West, Generating trees and the Catalan and Schröder numbers, *Discrete Math.* **146}** (1995), 247–262.

Department of Mathematics 2-375
Massachusetts Institute of Technology
Cambridge, MA 02139
rstan@math.mit.edu

NOTES

Edited by **Jimmie D. Lawson**

Many Correct Digits of π, Revisited

Gert Almkvist

Let n be a positive integer divisible by 4. Then

$$1 - t^2 + t^4 - t^6 + \cdots - t^{n-2} = \frac{1 - t^n}{1 + t^2}.$$

Integrating from 0 to 1 we obtain

$$\frac{\pi}{4} = \arctan 1 = \int_0^1 \frac{1}{1+t^2}\,dt = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots - \frac{1}{n-1} + R_n$$

where

$$R_n = \int_0^1 \frac{t^n}{1+t^2}\,dt.$$

One easily estimates

$$\frac{1}{2(n+1)} \le R_n \le \frac{1}{n+1},$$

so if we try to compute π taking $n/2$ terms in Gregory's series

$$4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots - \frac{1}{n-1}\right)$$

then the error is of order $1/n$. This did not prevent R. D. North from computing (to 40 digits)

$$4\left(1 - \frac{1}{3} + \frac{1}{5} - \cdots - \frac{1}{999999}\right)$$
$$= 3.14159\,\underline{0}6535\,89793\,2\overline{4}\underline{0}46\,26433\,832\underline{6}9\,50288\,4197$$

while $\pi = 3.14159\,26535\,89793\,23846\,26433\,83279\,50288\,4197.$

So only 4 out of 40 digits are wrong. This remarkable fact was explained in the celebrated paper **[2]** by the Borwein brothers and Karl Dilcher. Here we offer two other ways to do this.

First method. Make the substitution $t = e^{-x}$ in the remainder term

$$R_n = \int_0^1 \frac{t^n}{1+t^2}\,dt = \int_0^\infty \frac{e^{-(n+1)x}}{1+e^{-2x}}\,dx = \frac{1}{2}\int_0^\infty \frac{e^{-nx}}{\cosh x}\,dx$$

Let $f(x) = 1/\cosh x$. Then

$$f(x) = \sum_{k=0}^{m-1} E_{2k}\frac{x^{2k}}{(2k)!} + f^{(2m)}(\xi)\frac{x^{2m}}{(2m)!}$$

where the E_k's are the Euler numbers

$$E_0 = 1, E_2 = -1, E_4 = 5, E_6 = -61, \cdots$$

and $E_n = 0$ if n is odd.

Lemma. $|f^{(2m)}(x)| \le |E_{2m}|$ *for all* m *and* x.

Proof: (Jan Gustavsson): The trick is to use the fact that $f(x)$ is essentially its own Fourier transform

$$f(x) = \frac{1}{\cosh x} = \int_0^\infty \frac{\cos xt}{\cosh \pi t/2} dt.$$

Differentiate $2m$ times and take absolute values

$$|f^{(2m)}(x)| = \left| (-1)^m \int_0^\infty \frac{t^{2m} \cos xt}{\cosh \pi t/2} dt \right|$$

$$\le \int_0^\infty \frac{t^{2m}}{\cosh(\pi t/2)} dt = |f^{(2m)}(0)| = |E_{2m}|. \qquad \blacksquare$$

Integrating we obtain (using the Laplace transformation)

$$2R_n = \sum_{k=0}^{m-1} \frac{E_{2k}}{(2k)!} \int_0^\infty e^{-nx} x^{2k}\, dx + T_{(n,m)} = \sum_{k=0}^{m-1} \frac{E_{2k}}{(2k)!} \frac{(2k)!}{n^{2k+1}} + T_{(n,m)}$$

$$= \left(\frac{1}{n} - \frac{1}{n^3} + \frac{5}{n^5} - \frac{61}{n^7} + \cdots + \frac{E_{2m-2}}{n^{2m-1}} \right) + T_{(n,m)},$$

where

$$|T_{(n,m)}| \le \frac{|E_{2m}|}{n^{2m+1}}$$

by the Lemma. It follows that

$$4\left(1 - \frac{1}{3} + \frac{1}{5} - \cdots - \frac{1}{n-1}\right) = \pi - \frac{2}{n} + \frac{2}{n^3} - \frac{10}{n^5} + \frac{122}{n^7} - \cdots + 2T_{(n,m)}.$$

Putting $n = 10^6$ and $m = 3$ we find the digits different from those of π in North's computation.

Second method. Karl Dilcher told me how the paper with the Borweins was written during one hectic weekend. As it turns out, the main idea could have been found in about one second by Maple! Here is the crucial line

```
asympt(simplify((sum(-4*(-1)^j/(2*j-1), j=1..n/2)-Pi)/
                                          (-1)^(n/2+1)),n,8).
```

The answer is

$$-\frac{2}{n} + \frac{2}{n^3} - \frac{10}{n^5} + \frac{122}{n^7} + O\left(\frac{1}{n^9}\right).$$

The BBD-paper was cited by Andrews [1] as an example of the superiority of humans over the computer. Now the one-line Maple program above will of course be "gefundenes Fressen" for Zeilberger in his quest for arguments for the computer.

REFERENCES

1. G. E. Andrews, The death of proof? Semi-rigorous mathematics? You've got to be kidding!, *Math. Intelligencer* 16 (1994), No. 4, 16–18.
2. J. M. Borwein, P. B. Borwein, and K. Dilcher, Pi, Euler numbers, and asymptotic expansions, *Amer. Math. Monthly* 96 (1989), 681–687.
3. D. Zeilberger, Theorems for a price: Tomorrow's semi-rigorous mathematical culture, *Notices Amer. Math. Soc.* 40 (1993), 978–981.

Department of Mathematics
Lund University, Box 118
22100 Lund, Sweden
gert@maths. lth.se

A Note on a Cake Cutting Algorithm of Banach and Knaster

Martin L. Jones

The following famous problem was proposed by Steinhaus [4]. Can an object such as a cake be divided among n participants in such a way that each participant receives a piece equal to $1/n$ of the total cake according to his or her value system? Since its proposal there have been many solutions. One of the more elegant is the simple constructive algorithm attributed to Banach and Knaster by Steinhaus, which effects the desired division in the following manner. A long knife is passed parallel to itself slowly over the cake until some participant yells "stop," at which point the cake is cut and the piece just described by the knife is given to that participant. Ties are broken arbitrarily. The procedure is repeated with the remaining participants and with what remains of the cake. An implicit assumption of this algorithm is that the participant's values (measures) of the cake change continuously with the position of the knife. If the measures are assumed only to be nonatomic, that is, single points of cake have measure zero, then the Banach and Knaster algorithm might fail. Nonatomic measures can still have positive measure on lines or planes of cake. Imagine if a line of frosting parallel to the knife blade had measure one for each participant. Everyone would yell "stop" at the same time. In this case, a simple reorientation of the knife blade might solve the problem. That this can be done in general is the focus of this note.

Let μ be a probability measure defined on $(\mathbb{R}^m, \mathfrak{B}^m)$, where $\mathfrak{B}^m$ is the Borel σ-algebra in $\mathbb{R}^m$. The probability measure μ is said to be *nonatomic* if each individual m-tuple in $\mathfrak{B}^m$ has measure zero. However, subsets of dimension one or higher, such as lines and planes in $\mathfrak{B}^m$, may have positive μ measure. A set in $\mathfrak{B}^m$ that is a translation of a k-dimensional subspace of $\mathbb{R}^m$ will be referred to as an *affine subspace of dimension k*. For example, lines are affine subspaces of dimen-

sion 1, planes are affine subspaces of dimension 2, etc. For each $k = 1, 2, \ldots, m$, let

$$H_k := \{A \in \mathfrak{B}^m : A \text{ is an affine subspace of dimension } k\},$$

and let

$$H_k^* := \{A \in H_k : \mu(A) > 0 \text{ and } \mu(C) = 0$$
$$\text{whenever } C \subset A \text{ is an affine subspace with } \dim(C) < \dim(A)\}.$$

The set H_k^* is the collection of affine subspaces of dimension k that have positive μ measure, but all of whose *proper* affine subspaces do not. The following theorem restricts the cardinality of H_k^* under the assumption that μ is nonatomic.

Theorem. *Let μ be a nonatomic probability measure on* $(\mathbb{R}^m, \mathfrak{B}^m)$. *Then the set H_k^* is countable for each $k = 1, 2, \ldots, m$.*

Proof: Let $H_k^*(n) := \{A \in H_k^* : \mu(A) > 1/n\}$ for each $n \geq 1$, so $H_k^* = \bigcup_{n=1}^{\infty} H_k^*(n)$. We will show that card $H_k^*(n) \leq n$ by contradiction. Suppose card $H_k^*(n) > n$, and let $A_1, \ldots, A_{n+1}$ be distinct affine subspaces of dimension k in $H_k^*(n)$. By the principle of inclusion-exclusion we have

$$\mu\left(\bigcup_{j=1}^{n+1} A_j\right) = \sum_{j=1}^{n+1} \mu(A_j) - \sum_{i<j} \mu(A_i \cap A_j) + \cdots + (-1)^{n+2} \mu(A_1 \cap \ldots \cap A_{n+1}).$$

Since the intersection of any two or more distinct affine subspaces of dimension k is either empty or an affine subspace of dimension less than k, it follows from the nonatomic property of μ and the definition of H_k^* that

$$\mu\left(\bigcup_{j=1}^{n+1} A_j\right) = \sum_{j=1}^{n+1} \mu(A_j).$$

However, since each A_j belongs to $H_k^*(n)$ we have

$$\mu\left(\bigcup_{j=1}^{n+1} A_j\right) = \sum_{j=1}^{n+1} \mu(A_j) > \frac{n+1}{n} > 1,$$

contradicting the fact that μ is a probability measure. Therefore *card* $H_k^*(n) \leq n$, and *card* H_k^* is countable. ∎

Let the cake be represented by the solid unit sphere S in $\mathbb{R}^3$, let $\mathfrak{B}$ denote the Borel σ-algebra of subsets of S, and let μ be a nonatomic probability measure on $(S, \mathfrak{B})$. As the knife passes over the cake during the division process, the plane containing the knife blade partitions the cake into two pieces. To each possible orientation of the blade, there correspond two unit normals, $\mathbf{v}_1 = a_1\mathbf{i} + b_1\mathbf{j} + c_1\mathbf{k}$ and $\mathbf{v}_2 = a_2\mathbf{i} + b_2\mathbf{j} + c_2\mathbf{k}$. The points (a_1, b_1, c_1) and (a_2, b_2, c_2) on the surface of S are the points of tangency to the plane when it "enters" and "exits" the cake. Therefore, to each possible orientation of the knife blade there correspond two points on the surface of S. Let η be normalized Lebesgue measure on the surface

of S. The following corollary shows that the η measure of the "troublesome" orientations is zero.

Corollary 1. *Let μ be a nonatomic probability measure on $(S, \mathfrak{B})$. Let V be the set of all triples (a, b, c) such that $\mathbf{v} = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$ is a unit vector in $\mathbb{R}^3$ for which there exists a plane p in $\mathfrak{B}$ normal to $\mathbf{v}$ with $\mu(p) > 0$. Then $\eta(V) = 0$.*

Proof: First note that if p is a plane in $\mathfrak{B}$ with $\mu(p) > 0$, then either p contains a line with μ positive measure or it does not. Let V_1 be the set of triples $(a, b, c) \in V$ for which $\mathbf{v} = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$ is normal to a plane in $\mathfrak{B}$ that contains a line with positive μ measure. Let $V_2 = V \cap V_1^c$. To each $L \in H_1^*$ (so that $\mu(L) > 0$), there correspond uncountably many planes containing L. The normals to these planes all lie in the plane through the origin normal to L. Therefore, each $L \in H_1^*$ contributes a "great circle" of triples to V_1. By the Theorem, *card* H_1^* is countable, so V_1 consists of the triples on countably many great circles on the surface of S. If p does not contain a line with positive μ measure, but has positive μ measure itself, then $p \in H_2^*$. For each such p there are two unit normals. By the Theorem, *card* H_2^* is countable, so *card* V_2 is countable. Thus, $\eta(V) = \eta(V_1 \cup V_2) = \text{surface area } (V_1 \cup V_2) = 0$. ■

By repeating this argument for each of the n nonatomic probability measures, we obtain the following extension of Corollary 1.

Corollary 2. *Let $\mu_1, \ldots, \mu_n$ be nonatomic probability measures on $(S, \mathfrak{B})$. For each $i = 1, \ldots, n$, let V_i be the set of all triples (a, b, c) such that $\mathbf{v} = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$ is a unit vector in $\mathbb{R}^3$ for which there exists a plane p in $\mathfrak{B}$ normal to $\mathbf{v}$ with $\mu_i(p) > 0$. Then $\eta(\bigcup_{i=1}^n V_i) = 0$.*

Corollary 2 proves even more than was originally intended. Not only is it possible to utilize the Banach and Knaster algorithm for nonatomic measures, but if the orientation of the knife blade is chosen at random, a useable orientation will be chosen with probability one!

REFERENCES

1. S. J. Brams and A. D. Taylor, An envy-free cake division protocol. *Amer. Math. Monthly* 102 (1995), 9–18.
2. L. Dubins and E. Spanier, How to cut a cake fairly, *Amer. Math. Monthly* 68 (1961), 1–17.
3. T. P. Hill, Partitioning inequalities in probability and statistics, *IMS Lecture Notes* 22 (1993), 116–132.
4. H. Steinhaus, Sur la division progmatique, *Econometrika* (Supplement) 17 (1949), 315–319.
5. W. Stromquist, How to cut a cake fairly, *Amer. Math. Monthly* 87 (1980), 640–644.

Mathematics Department
University of Charleston, South Carolina
Charleston, South Carolina 29424
jonesm@math.cofc.edu

The Impossibility of Unstable, Globally Attracting Fixed Points for Continuous Mappings of the Line

Hassan Sedaghat

It is possible for a fixed point of a dynamical system to locally repel some trajectories, yet globally attract all trajectories. For example, consider the mapping

$$f_a(x) = \begin{cases} -2x & \text{if } x < a \\ 0 & \text{if } x \geq a \end{cases}$$

where a is any fixed positive real number. Then the first order difference equation

$$x_{n+1} = f_a(x_n) \qquad n = 0, 1, 2, 3, \ldots \tag{1}$$

has a solution

$$x_n = f_a^n(x_0) = \begin{cases} (-2)^n x_0 & \text{if } (-2)^{n-1} x_0 < a \\ 0 & \text{if } (-2)^{n-1} x_0 \geq a \end{cases}$$

for every choice of $x_0 \in \mathbf{R}$ (f_a^n represents the n-th iterate of f_a under function composition). Clearly, once $x_k \geq a$ for any k, then $x_n = 0$ for all $n \geq k$. In particular, every solution of (1) converges to zero, regardless of the choice of x_0. In this sense, the origin, which is the unique fixed point of f_a, is *globally attracting*. However, if $x_0 \neq 0$, then no matter how close x_0 is chosen to the origin, x_n must first exceed a before ultimately reaching the origin. Hence, the origin is *unstable* (in fact, locally repelling).

The preceding example shows that globally attracting fixed points that are *not* stable can easily occur in one-dimensional dynamical systems such as (1). Since f_a is discontinuous at $x = a$, it is natural to ask whether a *continuous* example of an unstable global point attractor can be constructed in one dimension. As the title of this note suggests, this is not possible. To see why continuous maps are nice in this sense, we need a local or asymptotic stability result from [**7**, p. 47]. Complete definitions of all concepts and terminology used here can be found in [**2**] and [**5**].

Criterion for asymptotic stability of fixed points: *A fixed point $\bar{x}$ of a continuous map f is asymptotically stable if and only if there is an open interval (a, b) containing $\bar{x}$ such that $f^2(x) > x$ for $a < x < \bar{x}$ and $f^2(x) < x$ for $\bar{x} < x < b$.*

The preceding criterion is remarkable for not requiring any differentiability conditions on f. Now we are ready to demonstrate our main result:

A continuous mapping of the real line cannot have an unstable fixed point that is globally attracting.

Suppose, on the contrary, that a continuous mapping f of the real line has an unstable fixed point $\bar{x}$ that is also globally attracting. Since there can be no periodic solutions, the iterate f^2 crosses the identity line only at $\bar{x}$. Hence, only

one of the following two cases is possible:

(I) $f^2(x) > x$ for $x < \bar{x}$ and $f^2(x) < x$ for $x > \bar{x}$;

(II) $f^2(x) < x$ for $x < \bar{x}$, or $f^2(x) > x$ for $x > \bar{x}$.

By the preceding Criterion, Case (I) implies stability and must therefore be ruled out; this leaves Case (II). Assume that $f^2(x) > x$ for $x > \bar{x}$, and let $x_0 > \bar{x}$. Then $f^2(x_0) > x_0$; as this implies $f^2(x_0) > \bar{x}$, repeated applications of f^2 to x_0 generate the increasing sequence

$$\bar{x} < x_0 < f^2(x_0) < f^4(x_0) < \cdots$$

By continuity, $f^{2n}(x_0) \to \infty$ as $n \to \infty$, implying that $\{f^n(x_0)\}$ does not converge to $\bar{x}$. The case $f^2(x) < x$ for $x < \bar{x}$ reaches a similar contradiction, so we conclude that our original assumption on $\bar{x}$ was false.

A natural question with regard to the preceding impossibility result is whether *one dimensionality* is necessary (in addition to continuity) in order to rule out the existence of unstable global point attractors. The answer is indeed affirmative, and examples of continuous (in fact, differentiable) planar maps having unstable, globally attracting fixed points exist in the literature; see, e.g., [**4**, p. 90], or the discretization of the continuous time example in [**1**, p. 59]. Unstable fixed points that are globally attracting can also arise in a continuous second order difference equation, which is a very special type of a two dimensional system. Generally, a second order difference equation has the form

$$y_{n+1} = F(y_n, y_{n-1}) \qquad n = 0, 1, 2, 3, \dots, \tag{2}$$

where $F\colon \mathbf{R}^2 \to \mathbf{R}$ and real numbers y_0, y_{-1} are specified as initial conditions. A *fixed point* of (2) is a solution of $F(y, y) = y$. The particular F that we discuss here is not an artificial construct of purely theoretical interest; rather, it comes from the classical Hicks model of the trade cycle, an early mathematical model that aimed to explain well-documented fluctuations in economic output or GNP that cause recessions periodically; see [**3**]. The simplified, static Hicks model with a single-period lag is given by equation (2) in which F is the continuous, piecewise linear mapping

$$F(u, v) = \min\{K, a + bu + c\max\{u - v, d\}\} \tag{3}$$

with constants $a, c > 0$, $d < 0$, $0 < b < 1$, and $K > a/(1-b)$. The ratio $a/(1-b)$ gives the unique fixed point (or *equilibrium*) $\bar{y}$ of the Hicks equation; for an explanation of the general Hicks model and the details of all derivations, see [**6**]. In particular, the negative number d is what Hicks calls the "floor level of induced investment." It is shown in [**6**] that, under these hypotheses, every non-equilibrium solution of (3) executes bounded, non-decaying oscillations about the unstable (and non-attracting) fixed point, as is expected of the "business cycle." But what happens when d approaches zero? It is in the limiting case $d = 0$ of the Hicks equation that the fixed point $\bar{y}$ turns into a global attractor, which is unstable if

$$c > \left(1 + \sqrt{1-b}\right)^2. \tag{4}$$

To see this, choose $y_{-1} = \bar{y}$ and $y_0 = \bar{y} + \varepsilon$, where $\varepsilon > 0$ is small enough so that $y_0 < K$. Then the trajectory $\{y_n\}$ develops according to the linear difference equation

$$y_{n+1} = a + (b + c)y_n - cy_{n-1},$$

which has exponentially divergent solutions, since condition (4) implies the exis-

tence of eigenvalues with magnitude greater than 1. Hence, $\bar{y}$ is unstable. Upon reaching K, however, the trajectory bounces down and obeys the first order equation

$$y_{n+1} = a + by_n,$$

whose solution clearly converges to $\bar{y}$. Generalizing this argument to arbitrary pairs of initial conditions is not hard, and establishes that $\bar{y}$ is globally attracting.

REFERENCES

1. N. P. Bhatia and G. P. Szego, *Stability Theory of Dynamical Systems*, Springer-Verlag, New York, 1970.
2. R. Devaney, *An Introduction to Chaotic Dynamical Systems*, 2nd ed., Addison-Wesley, Reading, MA, 1989.
3. J. R. Hicks, *A Contribution to the Theory of the Trade Cycle*, (2nd ed.) Clarendon Press, Oxford, 1965.
4. V. Lakshmikantham and D. Trigiante, *Theory of Difference Equations: Numerical Methods and Applications*, Academic Press, San Diego, 1988.
5. J. T. Sandefur, *Discrete Dynamical Systems: Theory and Applications*, Clarendon Press, Oxford, 1990.
6. H. Sedaghat, Bounded Oscillations in the Hicks Business Cycle Model and other Delay Equations, *J. Difference Equations Appl.* (to appear).
7. A. N. Sharkovsky, Yu.L. Maistrenko, and E.Yu. Romanenko, *Difference Equations and Their Applications*, Kluwer Academic Publishers, Dordrecht, 1993.

Department of Mathematical Sciences
Virginia Commonwealth University
Richmond, VA 23284-2014
hsedaghat@ruby.vcu.edu

[The try-works are] a place also for profound mathematical meditation. It was in the left-hand try-pot of the Pequod, with the soapstone diligently circling round me, that I was first indirectly struck by the remarkable fact, that in geometry all bodies gliding along the cycloid, my soapstone for example, will descend from any point in precisely the same time.

Herman Melville, *Moby Dick*, Chapter XCVI, The Try-Works
Contributed by Karl David, Wells College

UNSOLVED PROBLEMS

Edited by **Richard Nowakowski**

In this department the MONTHLY presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Nowakowski, Department of Mathematics & Statistics & Computing Science, Dalhousie University, Halifax NS, Canada B3H 3J5, rjn@cs.dal.ca

Divisors and Desires

Richard K. Guy

Zhang Ming-Zhi, Lin Da-Zheng, and Wang Weh-Hui noted that, if $\phi(n)$ and $\sigma(n)$ are the Euler totient function and the sum of divisors function, then it is well-known that

$\phi(n) + \sigma(n) = 2n$ is a necessary and sufficient condition for $n > 1$ to be prime,

and they asked if there were composite values of n for which $\phi(n) + \sigma(n) = kn$. For $k = 3$ they found 312, 560, 588, 1400, 85632, 147492, 556160, 569328, 1590816, 2013216, 3343776, 4695456, 9745728 and 12558912, while for $k = 4$ there are 23760, 59400, 153720, and 4563000.

Some questions that arise are:

1. Are there infinitely many solutions for each $k > 2$?
2. Is there an odd solution?
3. All these solutions are of shape $4m$. Is there a solution of shape $4m + 2$?

In partial answer to Questions 2 and 3, Zhang, Lin, and Wang show that an odd n must have at least 6 distinct prime factors, and that if $k \geq 4$ then n has at least 15 such. They also show, parallel to Euler's result on odd perfect numbers, that an odd solution is a perfect square and that a singly even solution must be of the form $n = 2p^a q^2$, where p is a prime with $p \equiv a \equiv 1 \bmod 4$ and $q \perp 2p$.

We read in [**1**, §B41] that if we average the σ- and ϕ-functions, and iterate, then since $\phi(n)$ is always even for $n > 2$ and $\sigma(n)$ is odd when n is a square or twice a square, we will sometimes reach a noninteger value. For example, $54, 69, 70, 84, 124, 142, 143, 144, 225\frac{1}{2}$; in this case we say that the sequence *fractures*. Another possibility, in view of the opening observation, is that such sequences can become constant, for example, $60, 92, 106, 107, 107, \ldots$.

4. Are there such iterated sequences that increase indefinitely without fracturing?

The arithmetic mean of these two functions is not always an integer. Their geometric mean is an integer for $n = 1, 14, 30, 51, 105, 170, \ldots$.

5. Are there infinitely many n such that the product $\phi(n)\sigma(n)$ is a perfect square?

The product is exactly divisible by n when $n = 1, 6, 18, 24, 28, 40, 54, 72, 84, 96, 117, 120, 135, 162, 196, 200, 216, 224, 234, 252, 270, 288, \dots$. The perfect and multi-perfect numbers are there, of course, but this doesn't necessarily imply that there are infinitely many. But if we note that the odd powers of the perfect numbers are also present, we see that there are.

6. What is the density of these numbers?

Whenever n is prime, then $\phi(n)\sigma(n)$ is one less than a square. This also happens for $n = 6, 22, 33, 44, 69, 76, 82, \dots$.

7. Are there infinitely many composite numbers of this kind?
8. The product $\phi(n)\sigma(n)$ is constant for $n = 55$, 56 and 57. Are there infinitely many such triples? Can there be longer runs?

Scott Forrest asks if, for every base b, there is a *d-number*, i.e., a number n the sum of whose digits, when written in base b, is equal to the number, $d(n)$, of its divisors. An example of a d-number in base 10 is 262144, since

$$d(262144) = d(2^{18}) = 19 = 2 + 6 + 2 + 1 + 4 + 4.$$

In fact they seem to be quite numerous, so we can ask more specifically:

8. For each base is there always a two-digit d-number?

The least ones for the first few bases are

$11_2 \quad 22_3 \quad 11_4 \quad 13_5 \quad 11_6 \quad 12_7 \quad 24_8 \quad 26_9 \quad 11_{10} \quad 13_{11} \quad 11_{12} \quad 15_{13} \quad 24_{14}$

For some bases the smallest 2-digit d-number requires a somewhat larger first digit. Examples are 88_{26} and 84_{124}; the largest that Forrest found with $b \le 2000$ is $10b + 2$ for $b = 1397$ where $d(13972) = d(2^2 \cdot 7 \cdot 499) = 12 = 10 + 2$.

9. Is the first digit of the smallest 2-digit d-number bounded independently of the base, or is it infinitely often greater than $\ln b$, for instance?

Forrest also found the following numbers of d-numbers less than b^5, i.e., with less than 6 base b digits:

b	2	3	4	5	6	7	8	9	10	11	12	13	14
#	7	26	63	125	220	588	563	1561	1533	2451	2095	6501	3487

The first few entries in the table are remarkably close to b^3, but probably this is a manifestation of the Strong Law of Small Numbers [**2**].

10. Are there infinitely many d-numbers for each base?
11. If $f(b, N)$ is the number of base b d-numbers less than N, can we find good upper and lower bounds for $f(b, N)$?

We are indebted to Andrew Bremner for confirming that the desires of the title are indeed base desires.

REFERENCES

1. Richard K. Guy, *Unsolved Problems in Number Theory*, 2nd edition, Springer-Verlag, 1994.
2. Richard K. Guy, The strong law of small numbers, this MONTHLY **95** (1988) 697–712.

PROBLEMS AND SOLUTIONS

Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**

with the collaboration of Paul T. Bateman, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Underwood Dudley, Michael A. Filaseta, Ira M. Gessel, Bart Goddard, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Robert Israel, Murray S. Klamkin, Daniel J. Kleitman, Fred Kochman, Frederick W. Luttmann, Frank B. Miles, M. J. Pelling, Richard Pfiefer, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before September 30, 1997; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.*

PROBLEMS

10585. *Proposed by Alta Kellogg, Ormond Beach, FL.* A sequence $a_0, a_1, a_2, \ldots$ of real numbers is called *strictly totally positive* (STP) if every submatrix of the Hankel matrix $(a_{i+j})_{i,j\geq 0}$ has positive determinant.

(a) Show that the sequence $C_0, C_1, C_2, \ldots$ of Catalan numbers, defined by $C_n = \frac{1}{n+1}\binom{2n}{n}$, is STP.

(b) Show that the sequence of Catalan numbers is minimal in the following sense: If $a_0, a_1, a_2, \ldots$ is an STP sequence of positive integers with $a_n \leq C_n$ for every n, then $a_n = C_n$ for every n.

10586. *Proposed by C. D. Aliprantis, Indiana University–Purdue University, Indianapolis, IN.* Let x and y be two nonnegative Lebesgue integrable functions on $[0, 1]$ satisfying

$$\int_0^1 e^{-tx(t)}\,dt \geq \int_0^1 e^{-ty(t)}\,dt.$$

Show that

$$\int_0^1 ty(t)e^{-tx(t)}\,dt \geq \int_0^1 tx(t)e^{-tx(t)}\,dt.$$

10587. *Proposed by Joaquin Gómez Rey, Madrid, Spain.* Let K_{2n} be the complete graph on $2n$ vertices. Let P_n be the probability that two random perfect matchings of K_{2n} are disjoint. What is $\lim_{n\to\infty} P_n$?

10588. *Proposed by Marcin Mazur, The University of Chicago, Chicago, IL.* Let $A_1A_2A_3$ be a triangle. For $i = 1, 2, 3$, let B_i be a point on side $A_{i+1}A_{i+2}$, where subscripts are taken modulo 3.

(a) Show that $|A_iB_{i+1}| + |B_iB_{i+1}| = |A_iB_{i+2}| + |B_iB_{i+2}|$ holds for $i = 1, 2, 3$ if and only if B_i is the midpoint of $A_{i+1}A_{i+2}$ for $i = 1, 2, 3$.

(b) Show that $|A_iB_{i+1}| + |A_iB_{i+2}| = |B_iB_{i+1}| + |B_iB_{i+2}|$ holds for $i = 1, 2, 3$ if and only if B_i is the midpoint of $A_{i+1}A_{i+2}$ for $i = 1, 2, 3$.

10589. *Proposed by Tim Keller, Fair Oaks, CA.* Fix $n \geq 3$ and let S be the set of positive integers congruent to 1 modulo n. A number $m \in S$ is called *indecomposable* if it is not the product of two smaller numbers in S. Problem 3 from the 1977 International Mathematical Olympiad asks for a number that can be expressed as the product of indecomposable numbers in more than one way. Show that the least such number is the product of two numbers each of the form $k(k+n)$.

10590. *Proposed by Robb Muirhead, University of Michigan, Ann Arbor, MI and Stephen Portnoy, University of Illinois, Urbana, IL.* Let X have a uniform distribution on the interval $[0, 1]$ and let $N_{m,k}$ be the digit in the mth place to the right of the decimal point in X^k.
(**a**) Find $\lim_{m\to\infty} P(N_{m,m} = i)$ for $i = 0, 1, 2, \ldots, 9$.
(**b**) Characterize those functions $k(m)$ for which $\lim_{m\to\infty} P(N_{m,k(m)} = i) = 1/10$ for $i = 0, 1, 2, \ldots, 9$.

10591. *Proposed by John Lomont, University of Arizona, Tucson, AZ.* Let $\{a_n\}_{n=1}^{\infty}$ be the sequence of real numbers defined by $x/\tanh^{-1}(x) = 1 - \sum_{n=1}^{\infty} a_n x^{2n}$.
(**a**) Prove that $\sum_{n=1}^{\infty} a_n = 1$.
(**b**) Prove that $a_n > 0$ for all $n \geq 1$.
(**c**) Prove that $a_1 > 3a_2 > 5a_3 > 7a_4 > 9a_5 > \cdots$.

SOLUTIONS

Small Sets Meeting All Circles

10345 [1993, 874]. *Proposed by George Baloglou, SUNY College at Oswego, Oswego, NY, and Fred Galvin, University of Kansas, Lawrence, KS.* Given a subset $\mathbf{X} \subset \mathbb{R}$ one obtains a subset $\mathbb{R}^2 \setminus \mathbf{X}^2$ of the plane by removing those points both of whose coordinates are in $\mathbf{X}$. If $\mathbf{X} \neq \mathbb{R}$, such a set always contains horizontal and vertical lines.
(**a**) Find such a set $\mathbf{X}$, of Lebesgue measure zero, for which $\mathbb{R}^2 \setminus \mathbf{X}^2$ contains no circle.
(**b**)* Is there such a set $\mathbf{X}$, of Lebesgue measure zero, for which every connected subset of $\mathbb{R}^2 \setminus \mathbf{X}^2$ consisting of more than one point contains a horizontal or vertical line segment?

Solution of (**a**) *by Paul J. Szeptycki, Ohio University, Athens, OH.* For any measure zero dense G_δ set $\mathbf{X} \subseteq \mathbb{R}$, $\mathbb{R}^2 \setminus \mathbf{X}^2$ contains no circles.

Let $C \subseteq \mathbb{R}^2$ be a circle. Let

$$C_1 = \{(x, y) \in C\colon\ x \in \mathbf{X}\} \text{ and } C_2 = \{(x, y) \in C\colon\ y \in \mathbf{X}\}.$$

Then both C_1 and C_2 are dense G_δ's relative to C. By the Baire Category Theorem, $C_1 \cap C_2 \neq \emptyset$, but $C_1 \cap C_2 \subseteq \mathbf{X}^2$.

Solution of (**b**) *by Randall Dougherty, Ohio State University, Columbus, OH.* The following theorem settles (**b**) negatively since every set of positive Lebesgue measure includes a perfect set. (A set is *perfect* if it is closed and has no isolated points.)

Theorem. *If* $\mathbf{X}$ *is a set of real numbers such that* $\mathbb{R} \setminus \mathbf{X}$ *includes a perfect set, then there is a Borel connected subset of* $\mathbb{R}^2 \setminus \mathbf{X}^2$ *consisting of more than one point that does not include a horizontal of vertical line segment.*

Proof. Let $\mathbf{P}$ be a perfect set included in $\mathbb{R} \setminus \mathbf{X}$; we may assume that $\mathbf{P}$ includes no interval (just reduce $\mathbf{P}$ if necessary). Fix a point z in $\mathbf{P}$, and let $\mathbf{D}$ be a countable dense subset of $\mathbf{P}$.

Let $\mathbf{B}_1$ be the subset of $\mathbf{P} \times \mathbb{R}$ consisting of those points (x, y) such that either $x \in \mathbf{D}$ and y is rational, or $x \in \mathbf{P} \setminus \mathbf{D}$ and y is irrational. This is the *Cantor tepee* space of Knaster and Kuratowski (Example 129 in L. A. Steen & J. A. Seebach, Jr., *Counterexamples in Topology*, Holt, Rinehart and Winston, 1970), with the apex moved up to infinity. As in the

cited reference, one can show that any two points of $\mathbf{B}_1$ on the same vertical line are in the same quasicomponent of $\mathbf{B}_1$; that is, no relatively open and closed subset of $\mathbf{B}_1$ contains one of the two points but not the other. Clearly $\mathbf{B}_1$ does not include any vertical line segment.

Let $\mathbf{B}_1' = \mathbf{B}_1 \cup \{(x, z)\colon\ x \in \mathbf{P}\}$. Then, since we have added at most one new point on each vertical line, $\mathbf{B}_1'$ does not contain any vertical line segment; and it is still the case that any two points of $\mathbf{B}_1'$ on the same vertical line are in the same quasicomponent of $\mathbf{B}_1'$, since (x, z) is a limit of points $(x, y) \in \mathbf{B}_1$.

Now let $\mathbf{B}_2' = \{(y, x)\colon\ (x, y) \in \mathbf{B}_1'\}$ (i.e., the reflection of $\mathbf{B}_1'$ through the diagonal $y = x$) and $\mathbf{B} = \mathbf{B}_1' \cup \mathbf{B}_2'$. Then $\mathbf{B}$ is a Borel set that is included in $(\mathbf{P} \times \mathbb{R}) \cup (\mathbb{R} \times \mathbf{P})$ and hence does not meet $\mathbf{X}^2$. If S is a vertical line segment then, since $\mathbf{P}$ is closed and includes no interval, there must be a subsegment S' of S that is disjoint from $\mathbb{R} \times \mathbf{P}$ and hence from $\mathbf{B}_2'$. Since S' is not included in $\mathbf{B}_1'$, S is not included in $\mathbf{B}$. A similar argument shows that $\mathbf{B}$ contains no horizontal line segments.

Finally, we show that $\mathbf{B}$ is connected by showing that every point of $\mathbf{B}$ is in the same quasicomponent as the point (z, z), since this implies that every relatively open and closed subset of $\mathbf{B}$ is either $\mathbf{B}$ or $\emptyset$, depending on whether or not it contains (z, z). If $(x, y) \in \mathbf{B}_1'$, then (x, y) and (x, z) are in the same quasicomponent of $\mathbf{B}_1'$ (and hence of $\mathbf{B}$), while (x, z) and (z, z) are in the same quasicomponent of $\mathbf{B}_2'$ (and hence of $\mathbf{B}$), so (x, y) and (z, z) are in the same quasicomponent of $\mathbf{B}$. The argument for $(x, y) \in \mathbf{B}_2'$ is similar. This completes the proof of the theorem.

Editorial comment. Both Randall Dougherty and the proposers gave constructions in part **(a)** showing that it is possible for X to be an F_σ set of measure zero (in particular, a meager set). Dougherty also gave the following result, which may be interpreted as saying that many sets satisfy a weaker form of the property in **(b)**. In particular, such sets would give an affirmative answer to a version of **(b)** in which "connected" is replaced by "path connected".

Theorem. *If $\mathbf{X}$ is a comeager subset of $\mathbb{R}$ (i.e., $\mathbb{R} \setminus \mathbf{X}$ is of the first category), then any closed connected subset of $\mathbb{R}^2 \setminus \mathbf{X}^2$ consisting of more than one point includes a horizontal or vertical line segment.*

The selected solvers solved both parts of the problem. In addition, part (a) was solved by A. N. 't Woord (The Netherlands) and the proposers.

Preserving Rationality without Being Completely Straight

10361 [1994, 175]. *Proposed by Emil A. Cornea, University of Bucharest, Bucharest, Romania, and Florin N. Diacu, University of Victoria, Victoria, B. C., Canada.* Do there exist nonlinear C^1 functions $f : \mathbb{R} \to \mathbb{R}$ such that for any rational x, $f(x)$ is also rational and for any irrational x, $f(x)$ is also irrational?

Editorial comment. Several readers noted that the function $f(x) = x\big(1 + |x|\big)^{-1}$ has the required property.

Most readers gave such examples, pieced together from fractional linear functions. Additional smoothness requires a different construction. Some readers used functions defined by an infinite process starting from an enumeration of the rationals. While such constructions are more complicated, they can be used to construct examples that have such additional properties as monotonicity, convexity, or real analyticity.

The ability to improve smoothness by using an inductive construction led Rick Mabry to question whether one could also destroy smoothness; that is, find an example in which f is a continuous nowhere differentiable function. Known examples of nowhere differentiable functions can be modified to take rationals to rationals, but it appears difficult to also assure that irrationals are taken to irrationals.

The following references to the examples in the literature were given: Rick Mabry cited F. S. Cater, Derivatives on countable dense sets, *Real Analysis Exchange* 11 (1985-6), 159–167; Frank Schmidt cited B. Neumann & R. Rado, Monotone functions mapping the set of rational numbers onto itself, *J. Australian Math. Soc.* 3 (1963), 282–287; and Dave L. Renfro cited Philip Franklin, Analytic transformations of everywhere dense point sets, *Trans. Amer. Math. Soc.* 27 (1925), 91–100 and W. D. Maurer, Conformal equivalence of countable dense sets, *Proc. Amer. Math. Soc.* 18 (1967), 269–270.

Solved by J. Alvarez (Spain), G. Bartoszek & W. Bartoszek (South Africa), P. Budney, D. Callan, M. W. Cook, R. Ehrenborg (Canada), D. Fung, K. P. Hart (The Netherlands), R. Holzsager, R. B. Israel (Canada), F.-A. Izadi (Iran), R. M. Lansangan, J. H. Lindsey II, O. P. Lossers (The Netherlands), R. Mabry, M. D. Meyerson, A. Nijenhuis, N. Passell, D. L. Renfro, K. Schilling, R. L. Schilling (Germany), F. Schmidt, G. L. Stanek, R. Tschiersch (Germany), R. B. Tucker, T. White, D. R. Witte & J. C. Lagarias, A. N. 't Woord (The Netherlands), and the NSA Problems Group.

Constancy and Commensurability

10380 [1994, 363]. *Proposed by Michael Slater, University of Bristol, Bristol, England.* Suppose that $f_1, \ldots, f_n$ are continuous real periodic functions, and that $\sum_{i=1}^{n} f_i$ is a constant function, while no sum of fewer than n of the f_i is a constant function. Show that the f_i have a common period.

Solution I by A. N. 't Woord, University of Technology, Eindhoven, The Netherlands. Let $c \in \mathbb{R}$ be the value of the constant function $\sum_{i=1}^{n} f_i$. Rearranging the f_i, we may assume that for $2 \le i \le r$, f_1 and f_i have a common period and that for $r < i \le n$, f_1 and f_i have no common period. Then $f_1, \ldots, f_r$ have a common period, so that $g = \sum_{i=1}^{r} f_i$ is periodic with period p, say. For $r < i \le n$, let p_i denote a period of f_i. Then $p/p_i \notin \mathbb{Q}$ for $r < i \le n$.

Let $x \in \mathbb{R}$. We show that $g(x) = c - \sum_{i=r+1}^{n} f_i(0)$. Let $\epsilon > 0$. Since the f_i are continuous, there exists $\delta > 0$ such that $|x - y| < \delta$ implies $|g(x) - g(y)| < \epsilon/n$ and $|y| < \delta$ implies $|f_i(0) - f_i(y)| < \epsilon/n$ for $r < i \le n$. Using Kronecker's approximation theorem, we find $k, k_i \in \mathbb{Z}$ and $y \in \mathbb{R}$ such that $|y - x - kp| < \delta$ and $|y - k_i p_i| < \delta$ for $r < i \le n$. Now

$$\begin{aligned}\left|g(x) - \left(c - \sum_{i=r+1}^{n} f_i(0)\right)\right| &= \left|g(x) - c + \sum_{i=r+1}^{n} f_i(0) - g(y) + c - \sum_{i=r+1}^{n} f_i(y)\right| \\ &\le \left|g(x) - g(y)\right| + \sum_{i=r+1}^{n} \left|f_i(0) - f_i(y)\right| \\ &= \left|g(x) - g(y - kp)\right| + \sum_{i=r+1}^{n} \left|f_i(0) - f_i(y - k_i p_i)\right| < \epsilon.\end{aligned}$$

Hence $g(x) = c - \sum_{i=r+1}^{n} f_i(0)$, so that $g = \sum_{i=1}^{r} f_i$ is a constant function. Therefore $r = n$ and we conclude that the f_i have a common period.

Solution II by Richard Holzsager, The American University, Washington, DC. Being periodic, the functions f_i are also almost periodic, so we may use the inner product

$$\langle f, g\rangle = \lim_{t\to\infty} \frac{1}{2t} \int_{-t}^{t} f(x)g(x)\,dx.$$

For any f, $\langle 1, f\rangle$ is the average value of f. Replacing each f_i by $f_i - \langle 1, f_i\rangle$ allows us to assume that each f_i has zero average, without changing the outcome of the problem. Assuming that none of the f_i is constant, $\{f_1, \ldots, f_n\}$ decomposes into equivalence classes consisting of functions sharing a common period. If we replace each equivalence class by its sum, we get a new family with the same properties, but with no two having the same period.

We show that if f and g do not have a common period (i.e., their periods are incommensurate), then $\langle f, g\rangle = \langle f, 1\rangle\langle 1, g\rangle$. The result follows easily from this, because it implies that $\langle f_i, f_j\rangle = 0$ for $i \neq j$, so $0 = \langle f_i, 0\rangle = \langle f_i, \sum f_j\rangle = \langle f_i, f_i\rangle$. But this is a contradiction since f_i is not identically zero.

Let a and b be the respective periods of f and g, with a and b incommensurate. For any t, let $f^t(x) = f(x+t)$. Then $\langle f, g\rangle = \langle f^{ma}, g^{nb}\rangle = \langle f^{ma-nb}, g\rangle$, the first equality holding because the functions are unchanged and the second because the inner product is translation invariant. Since a and b are incommensurate, any t is the limit of numbers of the form $ma - nb$, and since periodic functions are uniformly continuous, it follows that f^t is the uniform limit of functions of the form f^{ma-nb}, so $\langle f, g\rangle = \langle f^t, g\rangle$ for all t. Then $a\langle f, g\rangle = \int_0^a \langle f^t, g\rangle = \langle \int_0^a f^t, g\rangle = \langle a\langle f, 1\rangle, g\rangle = a\langle f, 1\rangle\langle 1, g\rangle$. Cancelling the factor a gives the result we need.

Editorial comment. An account of the properties of the inner product used in Solution II can be found in texts on harmonic analysis (e.g., see section VI.5 of Y. Katznelson, *An Introduction to Harmonic Analysis*, Dover, 1976). The example, $f_1(x) = \sin x/3 - \sin x/5$, $f_2(x) = \sin x/5 - \sin x/7$, and $f_3(x) = \sin x/7 - \sin x/3$, shows that the common period (210π in this example) need not be the *minimal* period of any of the f_i.

Solved also by D. Beckwith, P. Budney, R. J. Chapman (U. K.), M. Golomb, L. E. Mattics, E. Wolf, and the proposer.

Orthogonally Additive Functions on an Integer Lattice

10381 [1994, 363]. *Proposed by Marcin E. Kuczma, University of Warsaw, Warszawa, Poland.* Determine all real valued functions f on the integer lattice $\mathbb{Z}^2$ such that $f(\mathbf{u}+\mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v})$ for every pair of orthogonal vectors $\mathbf{u}$, $\mathbf{v}$ in $\mathbb{Z}^2$.

Solution I by O. P. Lossers, University of Technology, Eindhoven, The Netherlands. It is clear that the set of functions f with this property forms a (real) vector space, V say, and that $f(\mathbf{0}) = 0$. We show that a basis for this vector space is provided by the following functions: $f_1(x, y) = x$, $f_2(x, y) = y$, $f_3(x, y) = x^2 + y^2$, and

$$f_4(x, y) = \begin{cases} 1 & \text{if } x \text{ is odd and } y \text{ is even;} \\ -1 & \text{if } x \text{ is even and } y \text{ is odd;} \\ 0 & \text{if } x \text{ and } y \text{ have the same parity.} \end{cases}$$

That these functions satisfy the condition is trivial for f_1 and f_2. For f_3, it follows from the Pythagorean theorem. For f_4, one verifies directly that $f_4(x_1, y_1) + f_4(x_2, y_2) = f_4(x_1 + x_2, y_1 + y_2)$ for all pairs (x_1, y_1), (x_2, y_2) except those where either x_1 and x_2 are both odd, or y_1 and y_2 are both odd (but not both). Since the exceptional pairs can never be orthogonal, it follows that $f_4 \in V$.

Now the functions f_1, f_2, f_3, and f_4 are clearly linearly independent, so it suffices to show that V is 4-dimensional. This follows from the fact that any $f \in V$ is completely determined from its values on $(1, 0)$, $(-1, 0)$, $(0, 1)$, and $(0, -1)$. Indeed, we have $f(x, y) = f(x, 0) + f(0, y)$, $f(2x, 0) = f(x, x) + f(x, -x)$, and $f(2x+1, 0) = f(2x+1, 1) - f(0, 1)$, while $f(2x+1, 1) = f(x+1, x+1) + f(x, -x)$. These equations (and similar ones for $f(0, 2y)$, $f(0, 2y+1)$ and $f(1, 2y+1)$) show that, for any pair (x, y) with $|x| + |y| > 1$, $f(x, y)$ can be expressed as a function of "smaller" (x, y)'s.

Solution II, generalizing to $n \geq 3$, by Robert Patenaude, College of the Canyons, Valencia, CA. When the domain of f is $\mathbb{Z}^n$ for $n \geq 3$, the condition $f(\mathbf{u} + \mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v})$ for orthogonal $\mathbf{u}$ and $\mathbf{v}$ becomes more restrictive. Let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ be the unit coordinate vectors in $\mathbb{Z}^n$ and let $f(\mathbf{e}_k) = a_k$ and $f(-\mathbf{e}_k) = b_k$ for $k = 1, \ldots, n$. Applying the method of Solution I in the plane spanned by $\mathbf{e}_1$ and $\mathbf{e}_3$ gives $f(2\mathbf{e}_3) = a_1 + b_1 + 2a_3$, while use of the plane spanned by $\mathbf{e}_2$ and $\mathbf{e}_3$ gives $f(2\mathbf{e}_3) = a_2 + b_2 + 2a_3$. Hence $a_1 + b_1 = a_2 + b_2$. By similar reasoning, $a_1 + b_1, \ldots, a_n + b_n$ all have a common value,

say $2m$. For each index k and any integer u_k then, $f(u_k\mathbf{e}_k) = mu_k^2 - mu_k + a_ku_k$. In general, $f(\mathbf{u}) = f(\sum u_k\mathbf{e}_k) = \sum f(u_k\mathbf{e}_k) = m\sum(u_k^2 - u_k) + \sum a_ku_k$, or in terms of the obvious column vectors, $f(\mathbf{u}) = m(\mathbf{u}-\mathbf{1})^{\mathrm{T}}\mathbf{u} + \mathbf{a}^{\mathrm{T}}\mathbf{u}$. Conversely, for all such functions f, $f(\mathbf{u}+\mathbf{v}) - f(\mathbf{u}) - f(\mathbf{v}) = m(\mathbf{u}^{\mathrm{T}}\mathbf{v} + \mathbf{v}^{\mathrm{T}}\mathbf{u}) = 2m(\mathbf{u}\cdot\mathbf{v}) = 0$ for orthogonal $\mathbf{u}$ and $\mathbf{v}$, as required.

Editorial comment. The only property of $\mathbb{R}$ required for the statement of the problem is that it have an operation denoted $+$. Thus, Aaron Meyerowitz and Fred Richman considered functions from $\mathbb{Z}^n$ to an abelian group A. The space $F_n(A)$ of allowable functions then forms an abelian group under pointwise addition. The case of $A = \mathbb{Z}$ is paramount. The selected solutions are easily modified to obtain functions generating $F_n(\mathbb{Z})$ as a free abelian group. For general A, $F_n(A)$ consists of combinations of these functions with coefficients in A.

Similar spaces of functions defined on a vector space with an orthogonality relation have been studied. Also, if A allows multiplication by $1/2$, $F_n(A)$ will be the direct sum of an even part (functions satisfying $f(-\mathbf{u}) = f(\mathbf{u})$) and an odd part (functions satisfying $f(-\mathbf{u}) = -f(\mathbf{u})$). See J. Rätz, On orthogonally additive mappings, II, *Publ. Math. Debrecen* 35 (1988), 241–249 and J. Rätz and Gy. Szabó, On orthogonally additive mappings, IV, *Aequationes Math,* 38 (1989), 73–85 for related results from that theory.

Solved also by D. Callan, R. J. Chapman (U. K.), M. Golomb, R. Holzsager, N. Komanda, J. H. Lindsey II, S. C. Locke, A. D. Meyerowitz & F. Richman, N. Passell, J. Rätz (Switzerland), M. Reid, R. Richberg (Germany), R. M. Robinson, K. Schilling, J. Spencer, T. White, A. N. 't Woord (The Netherlands), and the proposer.

A Matrix of Tangents

10387* [1994, 474]. *Proposed by Stanley Rabinowitz, Westford, MA and Peter J. Costa, University of St. Thomas, St. Paul, MN.* Let $T_n = (t_{i,j})$ be the n-by-n matrix with $t_{i,j} = \tan(i+j-1)x$, i.e.,

$$T_n = \begin{pmatrix} \tan x & \tan 2x & \tan 3x & \cdots & \tan nx \\ \tan 2x & \tan 3x & \tan 4x & \cdots & \tan(n+1)x \\ \tan 3x & \tan 4x & \tan 5x & \cdots & \tan(n+2)x \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \tan nx & \tan(n+1)x & \tan(n+2)x & \cdots & \tan(2n-1)x \end{pmatrix}.$$

Computer experiments suggest that $\det(T_n)$ equals

$$(-1)^{\lfloor n/2 \rfloor} \sec^n nx \prod_{r=1}^{n-1} \left(\sin^2(n-r)x \sec rx \sec(2n-r)x\right)^r \times \begin{cases} \sin n^2 x & \text{if } n \text{ odd,} \\ \cos n^2 x & \text{if } n \text{ even.} \end{cases}$$

Prove or disprove this conjecture.

Solution by David P. Robbins, Center for Communications Research, Princeton, NJ. We evaluate a more general n-by-n determinant:

$$\det\left(\frac{ax_j + by_k}{x_j + y_k}\right) = \frac{(a-b)^{n-1}\left(a\prod x_j + (-1)^{n-1}b\prod y_k\right)\prod_{1\le j<k\le n}\left((x_j - x_k)(y_j - y_k)\right)}{\prod_{j,k=1}^n (x_j + y_k)}. \quad (1)$$

Setting $a = -i$, $b = i$, $x_j = e^{i(2j-1)\theta}$, and $y_j = e^{-i(2j-1)\theta}$ in (1) gives the desired formula.

To prove (1), note that the determinant has the form $P/\prod_{j,k}(x_j + y_k)$, where P is a polynomial that is homogeneous of degree n^2 in the x's and y's and homogeneous of degree n in a and b. Because the determinant vanishes when two of the x's or two of the y's are

equal, P is divisible by $\prod(x_j - x_k)\prod(y_j - y_k)$. If we set $a = b$, then all the entries become identical. Thus subtracting the first row from each of the other rows makes each of the other rows divisible by $(a - b)$, and P is divisible by $(a - b)^{n-1}$.

Degree considerations show that the remaining factor has the form $au + bv$, where u and v are homogeneous of degree n in the x's and y's. Setting $x_i = 0$ makes the first row of the determinant (and hence P) divisible by b; thus x_i divides u, and u is a constant multiple of $\prod x_i$. Similarly, v is a constant multiple of $\prod y_j$. Finally, since the determinant must remain the same if we exchange a with b and the x's with the y's, we may conclude that $v/\prod y_j = (-1)^{n-1}u/\prod x_i$.

Thus the left side of (1) is a constant multiple λ_n of the right side. To find λ_n, we take the special case $b = 0$. Then we may cancel $a^n x_1 \cdots x_n$ from both sides, obtaining

$$\det\left(\frac{1}{x_j + y_k}\right) = \lambda_n \frac{\prod_{1\le j<k\le n}((x_j - x_k)(y_j - y_k))}{\prod_{j,k=1}^{n}(x_j + y_k)}. \tag{2}$$

If we multiply both sides of (2) by x_1 and take the limits as $x_1 \to \infty$ and then $y_1 \to \infty$, we find that $\lambda_n = \lambda_{n-1}$. It is clear that $\lambda_1 = 1$, and the proof is complete.

The identity (2) (with $\lambda_n = 1$) appears in Thomas Muir, *A Treatise on the Theory of Determinants*, Dover, 1960, paragraph 353, page 348.

Editorial comment. H. van Haeringen proved another generalization, computing the determinant $D_n(a) = \det\left(\tan(a + i + j - 1)x\right)_{i,j=1}^{n}$, where a is any complex number. He showed that $D_{n-1}(a+1)D_{n+1}(a-1) = D_n(a+1)D_n(a-1) - (D_n(a))^2$ and also observed that $H_n(a) = \det\left((a + i + j - 1)^{-1}\right)_{i,j=1}^{n}$ satisfies this same recurrence.

Solved also by D. Callan, R. J. Chapman (U. K.), E. Fernández Moral (Spain), H. van Haeringen (The Netherlands), and D. J. Wright.

A Sure-Win Game of Solitaire

10390 [1994, 574]. *Proposed by Ognian Enchev, Boston University, Boston, MA.* A standard deck of 52 playing cards is arranged at random in 4 rows and 13 columns. Show that with finitely many transpositions of cards of the same value (e.g., 7♣ and 7♡, K♢ and K♠, and so on) all cards can be arranged in such a way that each column contains one club, one diamond, one heart, and one spade.

Solution I by John H. Smith, Boston College, Chestnut Hill, MA. Suppose that all suits occur in each of the first $k - 1$ columns, but that some suit y is missing from the kth column. By using at most k transpositions, we introduce suit y to the kth column without eliminating any suit from any of the first k columns.

Some suit x occurs at least twice in the kth column. Define a directed graph with 13 edges whose vertices are the columns, as follows. For each value i, draw an arrow from the column where value i occurs in suit x to the column where i occurs in suit y. Each of the first $k - 1$ columns has exactly one head and one tail, and column k has at least two tails and no head.

Since each of the first $k - 1$ columns has one head and one tail, a path that enters this set must leave it. This yields a path from column k that eventually reaches a column later than k (it may go there directly). Making the transpositions corresponding to edges on this path introduces suit y to column k as desired. Applying this argument at most three times brings all suits into column k. The last column comes for free. Thus we can solve the whole problem in at most $3(1 + 2 + \cdots + 11) = 198$ steps.

Solution II by Gerry Myerson, Macquarie University, Sydney, NSW, Australia. Let C_j be the set of values i such that a card of value i appears in column j. Since each set of k columns contains $4k$ cards, it contains cards of at least k values. Hence $\{C_1, \ldots, C_{13}\}$ satisfies Hall's

condition and has a system of distinct representatives, meaning that we can select cards of distinct values from the 13 columns. These cards have at least four cards in some suit; and using at most 9 transpositions, we can spread that suit over the 13 columns. Repeating the argument, we can correct a second suit using at most $13 - \lceil 13/3 \rceil = 8$ transpositions and a third suit using at most $13 - \lceil 13/2 \rceil = 6$ transpositions. The remaining suit is then also spread over the 13 columns, and we have solved the problem using at most 23 transpositions.

Editorial comment. If the initial configuration consists of 12 columns that have only one suit and a 13th column having all suits, then each of the 12 bad columns must be involved in at least 3 transpositions, which establishes a lower bound of 18 transpositions.

Hall's theorem and related results can be found in many books on combinatorics. In particular, it is Theorem 1.1 in chapter 5 of H. J. Ryser, *Combinatorial Mathematics*, Carus Mathematical Monographs, no. 14, MAA, 1963. Readers also were able to obtain concise proofs of the result by using more specific theorems discussed later in that chapter.

It is natural to generalize to m values and n suits. The method of Solution II solves the game in at most $mn - \sum_{k \leq n} \lceil m/k \rceil$ steps, and the example described at the beginning of these comments leads to a lower bound of $\lfloor m/n \rfloor n(n-1)/2$ steps.

Solved also by G. Anderson & C. Anderson, D. Beckwith, R. J. Chapman (U. K.), P. Griffin, H. Helfgott, R. Holzsager, N. Komanda, J. H. Lindsey II, J. H. van Lint (The Netherlands), S. C. Locke, O. P. Lossers (The Netherlands), K. D. McLenithan & K. D. McLenithan, D. K. Nester, R. E. Prather, R. M. Robinson, N. Shazeer, T. Tran, C. Vanden Eynden, T. White, A. N. 't Woord (The Netherlands), NSA Problems Group, Prague Problems Group (Czech Republic), WMC Problems group, and the proposer.

Asymptotic Solution of a Recurrence Relation

10403 [1994, 792]. *Proposed by David Doster, Choate Rosemary Hall, Wallingford, CT.* Define a sequence $\langle y \rangle$ recursively by $y_0 = 1$, $y_1 = 3$, and

$$y_{n+1} = (2n+3)y_n - 2ny_{n-1} + 8n$$

for $n \geq 1$. Find an asymptotic formula for y_n.

Solution by University of South Alabama Problem Group, Mobile, AL. Let $x_n = y_n + 2n + 1$. Then $x_0 = 2$, $x_1 = 6$, and

$$x_{n+1} - 2(n+1)x_n = x_n - 2nx_{n-1}$$

for $n \geq 1$. Hence $x_n - 2nx_{n-1} = 2$ for $n \geq 1$. Now let $z_n = x_n/(2^n n!)$. Then $z_0 = 2$ and $z_k - z_{k-1} = 2/(2^k k!)$ for $k \geq 1$. Summing this from $k = 1$ to $k = n$, we see that

$$z_n = 2\sum_{k=0}^{n} \frac{(1/2)^k}{k!} = 2\sqrt{e} - 2\sum_{k=n+1}^{\infty} \frac{(1/2)^k}{k!}.$$

Hence

$$y_n = 2^{n+1} n! \sqrt{e} - 2n - 1 - \frac{1}{n+1} - \frac{1}{2(n+1)(n+2)} - \frac{1}{4(n+1)(n+2)(n+3)} - \cdots.$$

Solved also by R. A. Agnew, E. S. Andersen & M. E. Larsen (Denmark), J. Anglesio (France), R. Bagby, R. Barbara (France), J. M. L. Besora (Spain), P. Bracken (Canada), A. Brown (Australia), M. Burger (Austria), N. P. Bhatia & W. O. Egerland, L. Cagliero & J. Lauret (Argentina), R. J. Chapman (U. K.), B. W. Conolly (U. K.), D. A. Darling, J. Davis, P. K. Desikan, R. S. Gautam (India), C. Georghiou (Greece), P. Griffin, J.-P. Grivaux (France), R. Holzsager, J. Howard, M. E. H. Ismail, P. G. Kirmser, B. G. Klein, K.-W. Lau (Hong Kong), G. N. Lewis, O. P. Lossers (The Netherlands), C. Mallinger (Austria), B. Margolis (France), H. Morris, I. Nemes (Austria), V. Novakov (Bulgaria), J. Ottenstein (Israel), M. A. Pinsky, H. Prodinger (Austria), M. Reid, N. C. Singer, A. Stenger, R. Stong, A. A. Tarabay (Lebanon), D. B. Tyler, J. H. van Lint (The Netherlands), M. Vowe (Switzerland), D. Zeitlin, NSA Problems Group, Prague Problems Group (Czech Republic), WMC Problems Group, and the proposer.

A Nonlinear Recurrence in Two Variables

10408 [1994, 793]. *Proposed by Peter W. Shor, AT&T Bell Labs, Murray Hill, NJ.* Suppose that a function f is defined as follows: $f(1,1)=1$, $f(i,0)=0$ for $i \geq 0$, $f(1,j)=0$ for $j \geq 2$, and

$$f(i,j)=3(2i-j-2)f(i-1,j)+(i-2j+3)f(i-1,j-1)$$

otherwise.
(**a**) Show that $\sum_{j=1}^{\lfloor i/2 \rfloor+1}(-1)^{j+1}f(i,j)=i(i+2)(i+4)\cdots(3i-4)$.
(**b**) Find a closed form expression for $f(2j-1,j)$.

Solution to (**b**) *by George E. Andrews, Penn State University, University Park, PA.* With the standard convention that $1/n!=0$ if n is a negative integer, we show that

$$f(i,j)=\begin{cases}0 & \text{if } i<2j-2,\\ \dfrac{3^{i+1-2j}2^{j-i}(i+j-1)(2i-j-1)!}{(i-2j+2)!(j-1)!} & \text{otherwise.}\end{cases}$$

The three boundary conditions are immediate. As for the recurrence,

$$\begin{aligned}&3(2i-j-2)f(i-1,j)+(i-2j+3)f(i-1,j-1)\\ &=\frac{3^{i+1-2j}2^{j-i}(2i-j-2)!}{(i-2j+2)!(j-1)!}\big(2(i+j-2)(i-2j+2)+3(i+j-3)(j-1)\big)\\ &=\frac{3^{i+1-2j}2^{j-i}(2i-j-2)!(2i-j-1)(i+j-1)}{(i-2j+2)!(j-1)!}=f(i,j).\end{aligned}$$

Thus the answer to (**b**) is $f(2j-1,j)=\dfrac{2^{1-j}(3j-2)(3j-3)!}{1!(j-1)!}=\dfrac{(3j-2)!}{2^{j-1}(j-1)!}$.

Solution to (**a**) *by National Security Agency Problems Group, Fort Meade, MD.* Define $g(i,j)=(-1)^{j+1}f(i,j)$. We start by proving the recurrence

$$\begin{aligned}3(9i^2-4)g(i,j)+27(2i+1)\big(g(i+1,j+2)-g(i+1,j+1)\big)&\\ +8g(i+2,j+1)-9g(i+2,j+2)&=0.\end{aligned}$$

Using the formula for $f(i,j)$ from part (**b**), we compute

$$8g(i+2,j+1)-9g(i+2,j+2)=\frac{(-1)^j(2i-j+1)!3^{i-2j+1}}{(i-2j+2)!(j+1)!2^{i-j}}A(i,j),$$

where

$$A(i,j)=4(2i-j+2)(i+j+2)(j+1)+(i+j+3)(i-2j+2)(i-2j+1).$$

Also

$$g(i+1,j+2)-g(i+1,j+1)=\frac{(-1)^{j+1}(2i-j-1)!3^{i-2j-2}}{(i-2j+1)!(j+1)!2^{i-j}}B(i,j),$$

where

$$B(i,j)=2(i-2j+1)(i-2j)(i+j+2)+9(2i-j)(j+1)(i+j+1).$$

Thus

$$\begin{aligned}&27(2i+1)(g(i-1,j+2)-g(i+1,j+1))+8g(i+2,j+1)-9g(i+2,j+2)\\ &=\frac{(-1)^j(2i-j-1)!3^{i-2j+1}}{(i-2j+2)!(j+1)!2^{i-j}}\big(A(i,j)(2i-j)(2i-j+1)\\ &\qquad -B(i,j)(2i+1)(i-2j+2)\big),\end{aligned}$$

which reduces to

$$= \frac{(-1)^j(2i-j-1)!3^{i-2j+1}}{(i-2j+2)!(j+1)!2^{i-j}}\big(3(3i-2)(3i+2)j(j+1)(i+j-1)\big)$$
$$= 3(9i^2-4)\frac{(-1)^j(2i-j-1)!(i+j-1)3^{i-2j+1}}{(i-2j+2)!(j-1)!2^{i-j}} = -3(9i^2-4)g(i,j),$$

by polynomial arithmetic.

We have shown that the recurrence holds within the region where $f(i,j)$ is nonzero. Using the formula for $f(i,j)$, it is straightforward to show that the recurrence also holds at both boundaries (provided that we extend f and g by defining $f(i,-1)=0$).

Now let $S(i)=\sum_{j=1}^{\lfloor i/2\rfloor+1} g(i,j)$. Summing the recurrence for $-1\le j\le \lfloor i/2\rfloor+1$ yields $3(9i^2-4)S(i)-S(i+2)=0$, and thus $S(i+2)=3(3i-2)(3i+2)S(i)$. The initial values $S(1)=f(1,1)=1$ and $S(2)=f(2,1)-f(2,2)=2$ agree with the result claimed. For $i>2$, we have

$$S(i)=(3i-8)(3i-4)3\cdot S(i-2)=\frac{S(i-2)}{i-2}(3i-4)(3i-6)(3i-8),$$

and the result follows by induction.

Both parts were solved by both solvers and by the proposer.

Multiple Floors and Ceilings

10414 [1994, 912]. *Proposed by R. J. Simpson, Curtin University of Technology, Perth, Australia, and W. F. Smyth, McMaster University, Hamilton, Ontario, Canada.* For a positive real number x, let $C(x)=\lceil x/\lceil\sqrt{x}\,\rceil\rceil+\lceil\sqrt{x}\,\rceil$ and, for $x\ge 1$, let $F(x)=\lfloor x/\lfloor\sqrt{x}\rfloor\rfloor+\lfloor\sqrt{x}\rfloor$.

(**a**) Express $C(x)$ in a form that requires only one evaluation of a square root.

(**b**) Express $F(x)$ in terms of $C(x)$.

Solution to (**a**) *by Allen Stenger, Gardena, CA.* We prove that $C(x)=\lceil 2\sqrt{\lceil x\rceil}\,\rceil$ for $x>0$. Let $n=\lceil\sqrt{x}\,\rceil$, which is fixed for $(n-1)^2<x\le n^2$. In this range, $C(x)=\lceil x/n\rceil+n$ is integer-valued and nondecreasing, starting at $2n-1$ and ending at $2n$. The switch occurs when $2n-1=x/n+n$, at $x=n(n-1)$. Thus $C(x)=2n-1$ for $(n-1)^2<x\le n(n-1)$ and $C(x)=2n$ for $n(n-1)<x\le n^2$.

Like $C(x)$, the formula $\lceil 2\sqrt{\lceil x\rceil}\,\rceil$ is integer-valued and nondecreasing for $(n-1)^2<x\le n^2$, and it can change values only at integers. At n^2-n, it equals $\left\lceil 2\sqrt{n^2-n}\right\rceil=\left\lceil\sqrt{(2n-1)^2-1}\right\rceil=2n-1$. At $n^2-n+1/2$, it equals $\left\lceil\sqrt{(2n-1)^2+3}\right\rceil=2n$. Thus the formula has the same value as $C(x)$ everywhere.

Solution to (**b**) *by Robert A. Agnew, FMC Corporation, Chicago, IL.* Suppose $x\ge 1$. Let $m=\lfloor\sqrt{x}\rfloor$, which is fixed for $m^2\le x<(m+1)^2$. In this range (arguing as above),

$$C(x)=\begin{cases}2m & \text{if } x=m^2\\ 2m+1 & \text{if } m^2<x\le m(m+1)\\ 2m+2 & \text{if } m(m+1)<x<(m+1)^2;\end{cases}\qquad F(x)=\begin{cases}2m & \text{if } m^2\le x<m(m+1)\\ 2m+1 & \text{if } m(m+1)\le x<m(m+2)\\ 2m+2 & \text{if } m(m+2)\le x<(m+1)^2.\end{cases}$$

It follows that $F(x)=C(x)-1$ unless $x\in\{m^2,m^2+m\}$ or $(m+1)^2-1\le x<(m+1)^2$, in which case $F(x)=C(x)$.

Editorial comment. Robert A. Agnew provided a simple formula for $F(x)$ analogous to that for $C(x)$: $F(x)=\lfloor 2\sqrt{\lfloor x\rfloor+1}\rfloor$. Other correct formulas for $C(x)$ include $\lceil 2\sqrt{\lceil 2x\rceil/2}\,\rceil$

(O. P. Lossers), $\left\lceil\sqrt{4\lceil x\rceil - 1}\right\rceil$ (Albert Nijenhuis), and $\left\lceil\sqrt{2(\lceil x^2/\lceil x\rceil\rceil + \lceil x\rceil)}\right\rceil$ (Western Maryland College Problems Group).

Solved also by D. Callan, J. Christopher, R. B. Eggleton, O. P. Lossers (Netherlands), D. K. Nester, A. Nijenhuis, A. A. Tarabay (Lebanon), NSA Problems Group, WMC Problems Group, and the proposers.

An Odd Integral Sum

10473 [1995, 745]. *Proposed by Emre Alkan (student), Bosphorus University, İstanbul, Turkey.* Prove that there are infinitely many positive integers m such that

$$\frac{1}{5\cdot 2^m}\sum_{k=0}^{m}\binom{2m+1}{2k}3^k$$

is an odd integer.

Solution by Ulrich Abel, Fachhochschule Giessen-Friedberg, Friedberg, Germany. Let $a_m = 2^{-m}\sum_{k=0}^{m}\binom{2m+1}{2k}3^k$. By the binomial formula,

$$a_m = \frac{\left(1+\sqrt{3}\right)^{2m+1} + \left(1-\sqrt{3}\right)^{2m+1}}{2^{m+1}} = \frac{1+\sqrt{3}}{2}\left(2+\sqrt{3}\right)^m + \frac{1-\sqrt{3}}{2}\left(2-\sqrt{3}\right)^m.$$

Thus the sequence satisfies a second order linear recurrence with characteristic roots $2\pm\sqrt{3}$. This yields $a_m - 4a_{m-1} + a_{m-2} = 0$ for $m \geq 2$. Using $a_0 = 1$ and $a_1 = 5$, we obtain the repeating sequence $1, 5, 9, 1, 5, 9, \ldots$ for the remainders modulo 10 of $\{a_m\}$. Thus each a_{3n+1} is an odd multiple of 5.

Editorial comment. The Anchorage Solutions Group found the same sum in the *College Mathematics Journal*, Problem #428 [1990, 246; 1991, 257].

Solved also by E. S. Andersen & M. E. Larsen (Denmark), J. Anglesio (France), D. Beckwith, K. L. Bernstein, J. C. Binz (Switzerland, with generalization), G. Bouza Allende (Cuba), P. Bracken, S. Byrd, D. Callan, R. J. Chapman (U. K.), T. H. Crocker, Q. H. Darwish (Oman), J. S. Frame (with generalization), Z. Franco, S. M. Gagola Jr. (with generalization), C. Georghiou (Greece), N. H. Guersenzvaig (Argentina), J. K. Haugland (Norway, with generalization), Th. Honold (Germany), W. Janous (Austria, with generalization), M. S. Klamkin (Canada), W. Koepf (Germany), R. A. Kopas, Y. H. Kwong, F. Lengyel (with generalization), C. Libis (three solutions), J. H. Lindsey II, J. H. van Lint (The Netherlands), O. P. Lossers (The Netherlands), T. Mack, P. McCartney, A. Nijenhuis, P. Paule (Austria), A. Pedersen (Denmark), C. Popescu (Belgium, with generalization), O. A. Saleh & T. J. Walters (with generalization), L. Scribani (South Africa, with generalization), R. P. Sealy (Canada), J. Seibert (Czech Republic), H.-J. Seiffert (Germany, with generalization), L. W. Shapiro & P. Peart, A. Sinefakopoulos (Greece), J. H. Steelman, A. Stenger, I. Vardi, M. Vowe (Switzerland), A. N. 't Woord (The Netherlands), Z. Wu, P. J. Zwier, Anchorage Solutions Group, GCHQ Problem Solving Group (U. K.), NSA Problems Group, Trinity University Problem Group, and the proposer.

A Telescoping Sum

10494 [1996, 74]. *Proposed by WMC Problems Group, Western Maryland College, Westminster, MD.* For each positive integer n, evaluate the sum

$$\sum_{k=0}^{2n}(-1)^k\binom{4n}{2k}\Big/\binom{2n}{k}.$$

Solution by Richard Holzsager, The American University, Washington, DC. The value is $-1/(2n-1)$. Let $\left[{n \atop k}\right] = \binom{2n}{2k}/\binom{n}{k}$. For $1 \leq k \leq n$, we compute

$$\begin{aligned}\left[{n \atop k}\right] + \left[{n \atop k-1}\right] &= \frac{\prod_{i=1}^{k}(2n-2i+1)}{\prod_{i=1}^{k}(2i-1)} + \frac{\prod_{i=1}^{k-1}(2n-2i+1)}{\prod_{i=1}^{k-1}(2i-1)}\\ &= \frac{2n}{2k-1}\left[{n \atop k-1}\right] = \frac{2n}{2n+1}\left[{n+1 \atop k}\right].\end{aligned}$$

Thus

$$\sum_{k=0}^{2n}(-1)^k\begin{bmatrix}2n\\k\end{bmatrix} = 1+1+\frac{4n-1}{4n-2}\sum_{k=1}^{2n-1}(-1)^k\left(\begin{bmatrix}2n-1\\k-1\end{bmatrix}+\begin{bmatrix}2n-1\\k\end{bmatrix}\right)$$
$$= 2-2\frac{4n-1}{4n-2} = -\frac{1}{2n-1}.$$

Editorial comment. Solvers used a variety of methods, including induction, the beta integral, Gauss's hypergeometric series summation, generating functions, and computer algebra. Several solvers noted that the identity appears as equation (4.26) on page 49 of H. W. Gould's *Combinatorial Identities* (Morgantown, WV, 1972).

Solved also by U. Abel (Germany), Z. Ahmed (India), E. S. Andersen & M. Larsen (Denmark), J. Anglesio (France), D. Beckwith, D. Bradley (Canada), J. T. Bruening, R. J. Chapman (U. K.), W. Chu & A. Marini (Italy), D. A. Darling, J. S. Frame, J. Grossman, M. Hoffman, A. Kaplan (France), C. Krattenthaler (Austria), J. H. Lindsey II, O. P. Lossers (The Netherlands), P. McCartney, R. Nelsen, J. H. Nieto, M. Petkovsek (Slovenia), H. Pollak, S. Radhakrishnan, O. A. Saleh & S. Byrd, L. Scribani (South Africa), H.-J. Seiffert (Germany), T. R. Shore, A. Sinefakopoulos (Greece), I. Sofair, J. H. Steelman, H. L. Stubbs, A. Tissier (France), T. V. Trif (Romania), R. B. Tucker, J. Van hamme (Belgium), M. Vowe (Switzerland), S. Wagon, H. S. Wilf, K. Williams (Canada), A. N. 't Woord (The Netherlands), Z. Wu, NSA Problems Group, USA Problems Group, and the proposer.

Boundedness Along Subsequences

10500 [1996, 75]. *Proposed by Jeffrey C. Lagarias and Peter W. Shor, AT&T Bell Laboratories, Murray Hill, NJ.* Consider the following three properties that a sequence $\{f(n) : n = 1, 2, \ldots\}$ of real numbers may have.

(P1) The sequence $\{f(n) : n = 1, 2, \ldots\}$ is bounded.

(P2) For each real $\lambda > 1$, the subsequence $\{f(\lfloor 2^{\lambda^n}\rfloor) : n = 1, 2, \ldots\}$ is bounded.

(P3) For each real $\lambda > 1$, the subsequence $\{f(\lfloor \lambda^{2^n}\rfloor) : n = 1, 2, \ldots\}$ is bounded.

Obviously (P1) $\Longrightarrow$ (P2) and (P1) $\Longrightarrow$ (P3). What other implications hold, if any?

Solution by Richard Holzsager, The American University, Washington, DC. Property (P2) is equivalent to (P1), but (P3) is strictly weaker.

To see the first fact, we need to show that if a function f is unbounded on the integers, then it is unbounded on the sequence $\lfloor 2^{\lambda^n}\rfloor$ for some $\lambda > 1$. Choose $1 < n_1 < n_2 < n_3 < \ldots$ so that $f(n_k)$ is strictly increasing to infinity. Let lg be the base 2 logarithm, and denote by I_k the image under lg lg of the interval $[n_k, n_k + 1]$. Note that each I_k has length less than 1. Starting with $J_1 = I_1$, inductively define a nested sequence $\{J_n\}$ as follows: Given $J_n = [a, b]$, let m be an integer large enough so that $m(b - a) > a + 1$. Choose $k = k_{n+1}$ so that $I_k = [c, d]$ satisfies $c > ma$. Enlarge m, if necessary, to m_{n+1} so that $m_{n+1}a < c \le (m_{n+1} + 1)a$. Then $d < c + 1 \le m_{n+1}a + a + 1 < m_{n+1}b$, so the interval $J_{n+1} = I_k/m_{n+1}$ is contained in the interior of J_n. The lengths of these intervals approach 0, so their intersection consists of a single point t. Then, for every n, the multiple mt is in the interior of I_k for $k = k_n, m = m_n$. This is equivalent to $n_k < 2^{2^{mt}} < n_k + 1$. If we choose $\lambda = 2^t$, then $\lfloor 2^{\lambda^m}\rfloor = n_k$, and f is unbounded on the sequence determined by λ.

For the second claim, let $f(n) = n$ if $n = 2^{2^k} - 1$, and 0 otherwise. By adding or deleting a finite number of terms of any sequence $\{\lfloor \lambda^{2^k}\rfloor\}$, we can assume $\sqrt{2} \le \lambda < 2$. Then for all k, we have $2^{2^{k-1}} \le \lambda^{2^k} < 2^{2^k}$. Furthermore, the difference between the last two quantities clearly increases with k, and is eventually greater than 1. This means that $f(\lfloor \lambda^{2^k}\rfloor)$ is eventually 0 for any λ.

Solved also by J. Merickel, GCHQ Problem Solving Group (U. K.), and the proposers.

REVIEWS

Edited by **Underwood Dudley**

Mathematics Department, De Pauw University, Greencastle, IN 46135

Table of Integrals, Series and Products: CD-ROM Version 1.0. By I. S. Gradshteyn and I. M. Ryzhik, edited by Alan Jeffrey. Academic Press, 1996, $79.95

Reviewed by **Jet Wimp**

I *adore* GRJ (Gradshteyn, Ryzhik, and Jeffrey), all 1200 pages of it. I have two copies: one for the office, one for home. If my upper body musculature were better developed, I would carry it to Europe with me, to Asia, to the Himalayas. Though many worthwhile identities are missing from GRJ, for instance those relating to differential and difference properties of hypergeometric functions, it can serve as a near equivalent to the combined five volumes in the Erdélyi set **[1]**, **[2]**. It's the ideal desert island book, though of course one has to get it there, and it may not be easy explaining to one's ex-shipmates why this hefty and venerable tome should be given preference over that moldy goatskin flask of water when planning the contents of the lifeboat.

Let me talk a bit about this volume, the book, not the CD I'm reviewing. Though called, modestly, *Table of Integrals, Series, and Products*, that doesn't begin to do justice to its contents. The appellation "sums" covers closed-form expressions of frequently occurring sums along with much material on the convergence of infinite series. These topics, and products, functional series, asymptotic series, and formulas from differential calculus comprise the introductory chapter, the 0th chapter of the volume. Also materializing in this chapter are some of those hoary, cabalistic functions that used to inhabit our mathematical books but no longer do: the Gudermannian, Lobachevsky's "angle of parallelism."

The first chapter is a treatment of elementary functions so comprehensive it will allow the owner to ditch all those tattered trig tables. The second, third, and fourth chapters deal with definite and indefinite integrals of elementary functions, and the fifth, sixth, and seventh deal with definite and indefinite integrals of special functions. Of course, to make the book self-contained, a discussion of special functions is required, and this the book has, in spades. Its eighth and ninth chapters contain a wonderful 200 page treatment of all the standard higher functions, and many of the results given there are absent from the Erdélyi volumes —indeed, absent from any mathematical treatment commonly available. The formulas are a lot of fun to read and one can, Jeopardy-like, shield from view the left hand side of an equation and ask what the right hand side represents, for instance:

ANSWER:

The series

$$\sum_{k=0}^{\infty} c_k z^k, \quad |z| < 2$$

where

$$c_{n+1} = \frac{\sum_{k=0}^{n} (-1)^{k+1} s_{k+1} c_{n-k}}{n+1}, \; n = 0, 1, 2, \cdots,$$

with $c_0 = 1, s_1 = \gamma = .57721\ldots, s_n = \zeta(n), n > 2.$

QUESTION: What is $\Gamma(z+1)$?

There are several supplementary chapters that practicing mathematicians will consider pure gold: a chapter on vector field theory, one on algebraic inequalities, one on integral inequalities, a chapter on matrices, one on determinants, one on norms, one on ordinary differential equations, and, to cap it off, a chapter containing Mellin, Fourier, and Laplace transforms; this chapter is vestigial, though, and doesn't offer a viable alternative to the standard tables **[4]**, **[5]**, **[6]**.

The earlier editions had gobs of mistakes, but thanks to dedicated readers who have filed their suggested emendations with the editor, with the fifth edition most of the mistakes have been weeded out. I found one though (I almost had to justify my credentials for writing this review). The power β on the right hand side of the transformation formula for the Appell function F_4, formula 3 on page 1083, should be $-\beta$.

Where, in its comprehensiveness, does this volume stand in comparison to other tables? Well, it far surpasses the quaint handwritten two-volume 1965 table of integrals of Gröbner and Hofreiter **[3]**. The transform sections, as pointed out, suffer in comparison to other tables. It certainly contains far less material than the mammoth five-volume set by Prudnikov and others with its 3500 pages of material **[7]**. The scope of **[7]**, though, probably exceeds what anyone will ever require. The reader may know of the Borges short story, "The Library of Babel" in which the author envisions a library whose first room contains a hundred or so books, the first containing a single page with the single letter "A," the second a page with the single letter "B," etc. The second room of the library starts with a book containing a single page with the letters "AA," followed by a book containing "AB, ..." I'm sure the reader gets the idea. Any desired text you want will be somewhere in that library; the problem is only one of information retrieval. The Prudnikov volumes come close to being a mathematical equivalent to Borges' imaginary construct.

And the *presentation* of the GRJ book: the binding, the appearance, the typography are all splendid. Remember the shabby photo offset reproductions of Russian books that held us in hostage twenty or so years ago—the malodorous books on oatmeal colored paper with the English intertext and the Cyrillic formulas? Nothing could be aesthetically more distant from these books than this. Academic Press has compiled a gorgeous volume. Naturally, I can think of things that should have been included that weren't, but when I am searching for a formula vital to some research objective, it's surprising how often it can be found in GRJ.

So what more remains to be said? Well, there's the old joke about the play heavily revised in tryouts on Philadelphia and Boston, and the producer saying to the playwright after a disastrous Broadway opening night, "You just died from improvement." Could this CD sound the death knell of a wonderful publishing concept?

On the drawing board, it must have seemed like a marvelous idea. A hugely popular enchiridion: let's make it available to anyone at the flick of a computer key. Let's make a CD out of it! Academic Press enlisted the talents of Lightbinders of San Francisco, who produced the CD using the opulent and flexible text display software called DynaText 2.3. I was impressed with the software, which might be a wonderful way to render some books computer accessible. But here, no. What went wrong? The very worst thing that could go wrong. *You can't see the formulas!* They are tiny, tiny, tiny. The linear notes for the CD cautioned that the reader should have Adobe File Manager to properly view the formulas. I ordered it. It helped not at all. One can do a mouse click in the upper right hand corner and things enlarge a bit, but not enough.

I noticed how Academic, perhaps reacting to an increasingly litigious society, had included in the flyleaf of the original book a warning:

> Academic Press and the editor have expended great effort to ensure that the material presented in this volume is correct.... However...neither Academic Press nor the editor shall be responsible for any errors, omissions, or damages arising from the use of this volume.

If this warning was ever warranted, it was in the liner notes of this CD. Damages, indeed. I estimate the chances of retrieving a correct formula from the screen display at about one out of three, that is, for anyone lacking microfocal vision. Now I was doing all this on a Mac platform: maybe someone's DOS or Windows equation was lushly readable, elegant, utile. If so, I offer my humblest apologies to Academic; my hearty congratulations go to any such customers. They may imagine themselves fortunate but, computer karma being what it is, they'll soon enough be victims to some other piece of flawed software.

I phoned customer support at Lightbinders, and there followed one of those Kafkaesque conversations that seem to be so much a feature of the computer age. "You can display the TeX representation of an equation," the consultant pointed out. "Then you can paste the equation into a TeX document." But copying by highlighting the selection didn't work. "Hmmmm," the consultant said. "Do a copy from the file menu," he ordered. It didn't work. "HHMMMM," he said. "Do a copy by depressing the command and C keys." No luck. "HHHMMMMMMMM," he said. "*That's strange*." There I was, facing this travesty of the desired equation, staring at me in its TeX metamorphosed form. Has the reader ever tried to reconstruct an equation from its TeX representation? Don't. It's easier to *derive* the equation. One of my favorite equations, a darling double integral for an Appell function, had become a sixteen-line porridge of "{}}}{}}{"s and "//\/\/\/"s. One can print the whole page containing the equation in question, but before one would do this, one would want to know what the equation says. The pericopes on notations and index of special functions look like so much flypaper. From the customer rep I got no satisfaction, and I was reminded of those primitive tribes where half the members speak one language, the other half another.

What else could go wrong? Something else did. You can't *find* things. The list of contents is too abbreviated, consisting only of the chapter headings in the book; further details require clicking the mouse on the displayed headings. Once again, a

Catch-22. You have to know where something is to find it. Where should you look for the definition of Euler's constant? In the introduction? In the definition of elementary functions? In the integral tables? No, far away in Chapter 9, Special Functions, buried several layers deep. The table of contents of the book, the honest to god paper book, has to be visually scanned too, but it is all in front of you: nothing is buried under something else. After struggling with the CD-ROM, I clutched the book to my chest, praying it would never go away.

Also, the software may have corrupted my system file, entailing a trek to the computer services center four city blocks away and, after I discovered the computer hardware was intact, a trek back to reinstall the system software. However, I make no accusations. Those ethereal and subtle software incompatibilities may be the only true things of the spirit, the only mysteries, vouchsafed to a technocratic society barreling into the 21st century. Let us revere them.

I love books, but I'm not a Luddite. I embrace software that works. Macbeth talked about being yanked untimely from his mother's womb, and I think that software—market pressure, no doubt—is often yanked untimely from the designer's noggin. If you're planning on adding this CD to your library, beware. Be certain your computing platform accommodates it in a way that makes it legible, convenient, and system compatible. Take nothing for granted. A hasty purchase means a wracking day on the phone to some callow customer rep in some distant part of the country, and no concomitant satisfaction.

REFERENCES

1. Erdélyi, A., et al., *Higher Transcendental Functions*, 3 vols., McGraw-Hill, New York, 1953.
2. Erdélyi, A., et al., *Tables of Integral Transforms*, 2 vols., McGraw-Hill, New York, 1954.
3. Gröbner, W. and N. Hofreiter, *Integraltafel*: erster Teil, *Unbestimmte Integrale*; zweiter Teil, *Bestimmte Integrale*, Springer-Verlag, Wien, 1965.
4. Oberhettinger, F., *Tabellen zur Fourier Transformation*, Springer-Verlag, Berlin, 1957.
5. Oberhettinger, F., *Tables of Bessel Transforms*, Springer-Verlag, New York, 1972.
6. Oberhettinger, F. and L. Badii, *Tables of Laplace Transforms*, Springer-Verlag, New York, 1973.
7. Prudnikov, A. P., Yu.A. Brychkov, and O. I. Marichev, *Integrals and Series*, 5 vols., Gordon and Breach, New York, 1986.

Department of Mathematics and Computer Science
Drexel University
Philadelphia, PA 19104
jwimp@mcs.drexel.edu

The Parsimonious Universe. By Stefan Hildebrandt and Anthony Tromba. Springer-Verlag New York, Inc., 1996, 344 pp., $32.

Reviewed by **Frank Morgan**

We are just beginning to understand how geometry rules the universe. We know that a round soap bubble has found the least-area way to enclose a given volume of air, but we do not know for sure whether the familiar double soap bubble provides the least-area way to enclose and separate two given volumes of air, despite the

much-publicized recent computer breakthroughs on the special case of equal volumes by Hass, Hutchings, and Schlafly [**3**], [**4**], [**10**]. The 100-year-old Kelvin Conjecture on the least-area way to partition space into equal volumes was disproved in 1994 by the new, less symmetric conjectured optimal structure of Weaire and Phelan [**18**], [**14**], [**11**], [**9**], computed to be about 0.3% better by Brakke's Evolver (see Brakke's home page at http://www.susqu.edu/facstaff/b/brakke/default.htm). The analogous planar conjecture states that regular hexagons provide the least-perimeter way to partition the plane into unit areas. Even this planar conjecture remains open today, despite abundant mistaken announcements and proofs, from H. Weyl's *Symmetry* [**19**, p. 85] (originally published in 1952) to the book under review (p. 225).

At the International Conference on Differential Geometry at IMPA in Rio de Janeiro, Brazil, July, 1996, Bessa and Jorge announced a proof of the Calabi Conjecture, that there are no complete bounded minimal surfaces in R^3. Meanwhile D. Hoffman and others continue to produce beautiful new examples of unbounded complete minimal surfaces; in Brazil, M. Traizet [**20**] explained how to get the latest by desingularizing families of intersecting vertical planes.

The May, 1996 Euroconference on Foams in Arcachon, France brought together mathematicians, scientists, engineers, and others to discuss applications of soap bubbles and foams to cell structures, wine, bread, cleansers, fertilizers, shaving cream, oil wells, mines, acoustics, drug delivery, fire extinguishers, nuclear safety, and concrete.

Behind all the mathematics lie the human stories. Hass and Schlafly got their idea for a computer proof on the double bubble during a calm stretch between rapids while kayaking down the south fork of the American River in northern California. Their work depended on previous undergraduate research, as further developed by Hutchings, now a graduate student at Harvard. As for Lord Kelvin, we do not know exactly where he got his idea, but his niece reported [**7**, pp. 46–47]:

> When I arrived here yesterday, Uncle William [Lord Kelvin] and Aunt Fanny met me at the door, Uncle William armed with a vessel of soap and glycerine prepared for blowing soap-bubbles, on a tray with a number of mathematical figures made of wire. These he dips into the soap-mixture and a film forms or adheres to the wires very beautifully and perfectly regularly. With some scientific end in view he is studying these films.

Weaire and Phelan found their better structure in nature as a certain "clathrate." Brakke himself had searched long in vain, while all the time Linus Pauling's text, *The Nature of the Chemical Bond* [**13**], with a picture on page 471 of the new clathrate structure, sat on a shelf above his head.

The Parsimonious Universe by Hildebrandt and Tromba provides an exquisite liberal essay on the role of geometric minimization in the structure of the physical universe. Anyone familiar with the magnificent earlier version, published under the title *Mathematics and Optimal Form* (and reviewed by me in the Monthly [**12**]), will be amazed to find the new version a level still more attractive, with even more choice color illustrations, a still more appealing layout, and new material. Previously inaccessible documents from the former East Germany include an early letter of Leibniz on the principle of least action and the Copernicus portrait at Leipzig. In a virtually unknown 1630 engraving, Queen Dido directs her attendants to cut an ox hide into thin strips and arrange them into a large circular enclosure of land, to take fullest advantage of the terms of her bargain.

Hildebrandt recently mentioned to me the question they address of "whether Newton had originally written his *Principia* in the language of infinitesimal calculus or not. The editor of Newton's mathematical papers, Derek Whiteside, thinks the first, Michael Nauenberg from [UC] Santa Cruz, a well-known theoretical physicist..., is a recent strong proponent of the second opinion, and we have learned that an exciting discussion is taking place and might increase in the next years."

There is an illustrated account of H. Wente's 1984 discovery of a closed immersed constant-mean-curvature surface, although nothing of more recent examples of N. Kapouleas and others.

A chapter on Soap Films includes a nice account of Jean Taylor's mathematical derivation **[17]** of Plateau's laws **[15]**: that soap films meet only in threes at 120 degree angles along curves and that such curves meet only in fours at about 109 degree angles at points. There is an untold story here worth adding. The first step of the proof is to classify all nine nontrivial nets of geodesics on the sphere meeting in threes at 120 degree angles. The second step is to show that all but two of them do not produce stable soap film cones. Lamarle **[8]**, a contemporary mentioned by Plateau, took up the project but missed one of the nine cases. Alâdar Heppes, unaware of Lamarle's work, produced a complete proof but published only the first step **[5]**. Taylor **[17]**, initially unaware of Lamarle or Heppes, repeated the task, but ironically made a mistake on one of the nine cases. John Sullivan **[16]** explained a beautiful dual approach to the first step. Heppes finally wrote up the second step of his proof **[6]**. When writing the paper, he checked *Mathematical Reviews*, discovered Taylor's paper, and thence learned of Lamarle's work for the first time. In 1994, I asked an e-mail correspondent in Budapest of the same name, "You are not related to the Heppes who classified geodesic nets on the sphere, so central to the classification of soap film singularities, are you?" The reply: "The answer for your question is yes, I am related, and the relation is identity." He finally met Taylor, Sullivan, and me at a special session on Soap Bubble Geometry, which I organized at the 1995 Burlington Mathfest.

Although Hildebrandt and Tromba do not mention even the names of many contemporary scientists (a gap fortunately filled by the series on *Mathematical People* by Albers, Alexanderson, and Reid **[1]**, **[2]**), they give marvelous accounts of historical figures from Archimedes to Newton to Max Planck. Planck defended Leibniz's philosophy that our world is the best among all possible worlds in the face of world tragedy. The book ends with a touching postscript on Planck's own personal tragedies:

> Planck's oldest son Karl was killed at Verdun in 1916; both daughters died when giving birth to their first child; the younger son Erwin, because he had been involved in von Stauffenberg's assassination attempt of July 20, 1944, was sentenced to death by Hitler's criminal judge Freisler and was executed in January 1945. Shortly thereafter, Planck lost all his personal belongings, and at almost eighty-seven years of age he found himself in the great trek of refugees to the West, just like millions before and after him.

REFERENCES

1. Donald J. Albers and G. L. Alexanderson, eds., *Mathematical People*, Birkhäuser (Boston) in collaboration with the Mathematical Association of America, 1985.
2. Donald J. Albers, G. L. Alexanderson, and Constance Reid, eds., *More Mathematical People*, Academic Press, San Diego, 1990.
3. Joel Hass, Michael Hutchings, and Roger Schlafly, The double bubble conjecture, *Elec. Res. Ann. AMS* 1 (1995), 98–102.

4. Joel Hass and Roger Schlafly, Bubbles and double bubbles, *American Scientist*, Sept.–Oct., 1996, 462–467.
5. A. Heppes, Isogonale sphärischen Netze, *Ann. Univ. Sci. Budapest Eotvos Sect. Math.* 7 (1964), 41–48.
6. A. Heppes, On surface-minimizing polyhedral decompositions, *Disc. Comp. Geom.* 13 (1995), 529–539.
7. Agnes Gardner King, *Kelvin the Man, a Biographical sketch by his Niece*, Hodder and Stoughton Ltd., London, 1925.
8. Ernest Lamarle, Sur la stabilité des systèmes liquides en lames minces, *Mém. Acad. R. Belg.* 35 (1864), 3–104.
9. Frank Morgan, 100-year-old Kelvin Conjecture disproved by Weaire and Phelan, *Math Horizons*, to appear.
10. Frank Morgan, The double bubble conjecture, *MAA FOCUS*, December, 1995.
11. Frank Morgan, *Geometric Measure Theory: a Beginner's Guide*, Academic Press, 2nd edition, 1995.
12. Frank Morgan, Review of *Mathematics and Optimal Form* (Hildebrandt/Tromba), this *Monthly* 95 (1988), 569–575.
13. Linus Pauling, *The Nature of the Chemical Bond*, 3rd edition, Cornell Univ. Press, Ithaca, 1960.
14. Ivars Peterson, Constructing a stingy scaffolding for foam, *Science News*, March 5, 1994.
15. J. A. F. Plateau, *Statique Experimentale et Theorique des Liquides Soumis aux Seules Forces Moleculaires*, Paris, Gauthier-Villars, 1873.
16. John Sullivan, Convex deltatopes in all dimensions, and polyhedral soap films, preprint (1994).
17. Jean E. Taylor, The structure of singularities in soap-bubble-like and soap-film-like minimal surfaces, *Ann. Math.* 103 (1976), 489–539.
18. Denis Weaire and Robert Phelan, A counter-example to Kelvin's conjecture on minimal surfaces, *Phil. Mag. Lett.* 69 (1994), 107–110.
19. H. Weyl, *Symmetry*, Princeton University Press, 1989.
20. Martin Traizet, Construction de surfaces minimales en recollant des surfaces de Scherk, *C. R. Acad, Sci. Paris Ser. I Math.* 322 (1996), 451–453.

Department of Mathematics
Williams College
Williamstown, MA 01267
frank.morgan@williams.edu

Only he who knows what mathematics is, and what its function in our present civilization, can give sound advice for the improvement of our mathematical teaching.

Hermann Weyl, *Collected Works*, Volume I (opposite Weyl photograph)
Contributed by Hung-Hsi Wu, University of California at Berkeley

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

Mathematics Appreciation, S(13–14). *Short Stories From the History of Mathematics.* Robert E. Knauff. Carolina Mathematics (Carolina Biological Supply Co., 2700 York Road, Burlington, NC 27215), 1996, v + 99 pp, $10.50 (P). [ISBN 0-89278-435-0] 130 anecdotes, many well-known (Indiana legislature and π, origins of L'Hopital's rule, incidents from the lives of various mathematicians). DB

Precalculus, T(13: 1). *Precalculus: Enhanced with Graphing Utilities.* Michael Sullivan, Michael Sullivan, III. Prentice Hall, 1996, xxii + 1054 pp. [ISBN 0-02-343742-1] Traditional content, but emphasis and pedagogical approach influenced by technology. Includes open-ended problems, writing exercises, and collaborative projects.

Precalculus, T(13: 1). *Contemporary Precalculus: A Graphing Approach, Second Edition.* Thomas W. Hungerford. Saunders College, 1997, xviii + 871 pp, $45. [ISBN 0-03-018544-0] New edition reflects changes in available technology and user feedback. (*First Edition*, TR, April 1995.)

Precalculus, T(13: 1). *A Graphical Approach to Precalculus.* E. John Hornsby, Jr., Margaret L. Lial. HarperCollins, 1996, xxvi + 981 pp, $62.50. [ISBN 0-673-99966-1] Graphing calculator technology introduced in first chapter and used throughout. Includes group discussion activities and exercises "that make students aware of the connections between topics . . . ."

Education, P, L. *Windows on Mathematical Meanings: Learning Cultures and Computers.* Richard Noss, Celia Hoyles. Math. Educ. Lib., V. 17. Kluwer Academic, 1996, xi + 275 pp, $130. [ISBN 0-7923-4073-6] A comprehensive foundation for a revolution intended to make visible (and meaningful) the pervasive but "dead" mathematics "of the few," that is designed "to facilitate the priorities of those who control it" by making its power "invisible to the many." Rooted in Seymour Papert's vision of Logo as a children's mathematics, this analysis offers computers as the salvation of mathematics—as a tool to make our mathematical culture visible to the many. LAS

Education, P, L. *Assessment: Problems, Developments and Statistical Issues: A Volume of Expert Contributions.* Eds: Harvey Goldstein, Toby Lewis. Wiley, 1996, xvi + 273 pp, $74.95. [ISBN 0-471-95668-6] 15 chapters on all aspects of assessment: history, psychometric models, bias, comparison of institutions, educational testing, vocational assessment, international comparisons, and much more. Formula-free exposition broadens the potential audience to administrators and policy makers. LAS

Combinatorics, T(16–17: 1), L. *Introduction to Combinatorics.* Martin J. Erickson. Ser. in Disc. Math. & Optim. Wiley, 1996, xii + 195 pp, $59.95. [ISBN 0-471-15408-3] Main categories: existence (e.g., pigeonhole principle, partial orders, Ramsey theory), enumeration (including tableaux, Pólya theory), and construction (codes, designs). Open problems interspersed throughout text. LC

Discrete Mathematics, T(15–16: 1), L. *Elements of Algebraic Coding Theory.* L.R. Ver-

mani. Math. Ser. Chapman & Hall, 1996, viii + 254 pp, (P). [ISBN 0-412-57380-6] Primary focus is group/linear codes. Linear algebra prerequisite; abstract algebra concepts introduced as needed. Mathematically rigorous, but accessible to undergraduates. LC

Discrete Mathematics, T(13–14: 1). *Discrete Mathematics, Fourth Edition.* Richard Johnsonbaugh. Prentice Hall, 1997, xiv + 701 pp. [ISBN 0-13-518242-5] New to this edition: sections on binary and hexadecimal numbering systems, RSA public-key cryptosystems, more examples and exercises. (*Second Edition*, TR, May 1990.) TH

Number Theory, T(17–18: 1), S, P, L*. *Algorithmic Number Theory, Volume I: Efficient Algorithms.* Eric Bach, Jeffrey Shallit. Found. of Comp. MIT Pr, 1996, xvi + 512 pp, $55. [ISBN 0-262-02405-5] A thorough introduction to computational number theory. Covers GCD, calculations over finite rings and fields, primality testing. Many exercises. Extensive notes and bibliography. DB

Group Theory, P. *Modules and Group Algebras.* Jon F. Carlson. Lect. in Math. Birkhäuser Boston, 1996, xi + 91 pp, $26.50 (P). [ISBN 0-8176-5389-9]

Calculus, T(14: 1), S. *Functions of Two Variables.* Seán Dineen. Chapman & Hall, 1995, x + 189 pp, $23.95 (P). [ISBN 0-412-70760-8] A gentle introduction to multivariable calculus. Discusses maximum and minimum problems, plane curves, and integration theory on $\mathbb{R}^2$. Concise and informal exposition. AO

Real Analysis, S(17), P. *Advanced Analysis on the Real Line.* R. Kannan, Carole King Krueger. Universitext. Springer-Verlag, 1996, ix + 259 pp, $45 (P). [ISBN 0-387-94642-X] Presents topics not usually covered in textbooks on measure theory (e.g., approximate continuity, approximate differentiability, Hausdorff measure). Assumes knowledge of real analysis and some measure theory. PG

Differential Equations, T(16–18: 1), L. *Differential Equations and Dynamical Systems, Second Edition.* Lawrence Perko. Texts in Appl. Math., V. 7. Springer-Verlag, 1996, xiv + 519 pp, $49.95. [ISBN 0-387-94778-7] Nice text for a non-introductory course. Very clear coverage of linear systems. Discusses nonlinear systems (local dynamics, global, and bifurcations) at a level between Strogatz and Guckenheimer–Holmes. Many useful exercises and solved examples. (*First Edition*, TR, March 1992.) DK

Differential Equations, T(16–17: 3). *Multiple Scale and Singular Perturbation Methods.* J. Kevorkian, J.D. Cole. Appl. Math. Sci., V. 114. Springer-Verlag, 1996, viii + 632 pp, $59. [ISBN 0-387-94202-5] Revised and updated version (including new material) of the author's 1981 text *Perturbation Methods in Applied Mathematics* (TR, December 1981). Thorough survey of standard perturbation methods. Key techniques motivated by examples; theory presented in detail. Topics include limit process expansions, method of multiple scales, averaging transformations. MPR

Differential Equations, T(14).** *Elementary Differential Equations and Boundary Value Problems, Sixth Edition.* William E. Boyce, Richard C. DiPrima. Wiley, 1997, xvi + 749 pp. [ISBN 0-471-08955-9] Changes include a substantially revised chapter on numerical methods, more emphasis on investigating how a solution depends on a parameter, and a greater emphasis on visualization. (*Fifth Edition*, TR, January 1993.) AO

Partial Differential Equations, P. *Markov Processes and Differential Equations: Asymptotic Problems.* Mark Freidlin. Lect. in Math. Birkhäuser Boston, 1996, vi + 153 pp, $34.95 (P). [ISBN 0-8176-5392-9]

Partial Differential Equations, P. *Weak and Measure-valued Solutions to Evolutionary PDEs.* J. Málek, *et al.* Appl. Math. & Math. Comput., V. 13. Chapman & Hall, 1996, xi + 317 pp. [ISBN 0-412-57750-X]

Partial Differential Equations, P. *Minimax Theorems.* Michel Willem. Prog. in Nonlinear Diff. Eqts. & Their Applic., V. 24. Birkhäuser Boston, 1996, viii + 162 pp, $49.50. [ISBN 0-8176-3913-6]

Partial Differential Equations, P. *On Spectral Theory of Elliptic Operators.* Yuri Egorov, Vladimir Kondratiev. Oper. Theory: Adv. & Applic., V. 89. Birkhäuser Boston, 1996, x + 328 pp, $151.50. [ISBN 0-8176-5390-2]

Dynamical Systems, T(18: 3), P. *Hysteresis and Phase Transitions.* Martin Brokate, Jürgen Sprekels. Appl. Math. Sci., V. 121. Springer-Verlag, 1996, x + 357 pp, $59.95. [ISBN 0-387-94763-9] Hysteresis is the mathematical study of nonlinear "looping" phenomenon found in diverse areas such as ferro-magnetism, thermostat design, super-conductivity, and spin glasses. Detailed coverage of hysteresis from a variety of perspectives. Includes an introduction to the required mathematics. MPR

Dynamical Systems, P. *Dynamical Systems: Differential Equations, Maps and Chaotic Behaviour.* D.K. Arrowsmith, C.M. Place. Chap-

man & Hall, 1995, x + 330 pp, $57.95 (P). [ISBN 0-412-39080-9] Paperback republication of 1992 edition (TR, June–July 1994).

Numerical Analysis, C, P, L. *Numerical Algorithms with Fortran.* Gisela Engeln-Müllges, Frank Uhlig. Springer-Verlag, 1996, xxii + 602 pp, $49.95, with CD-ROM, [ISBN 0-540-60529-0]; *Numerical Algorithms with C*, xxii + 596 pp, $49.95, with CD-ROM. [ISBN 3-540-60530-4] Brief mathematical and computational descriptions of over 150 numerical algorithms together with programs implementing them. A translation, with a few updates, of the *Seventh Edition* (1993) of *Numerik-Algorithmen mit FORTRAN 77-Programmen.* AO

Numerical Analysis, T*(16–17: 2), L. *Numerical Analysis: Mathematics of Scientific Computing, Second Edition.* David Kincaid, Ward Cheney. Brooks/Cole, 1996, xii + 804 pp, $64. [ISBN 0-534-33892-5] A well-written text. Includes some non-traditional topics (e.g., the multigrid method, delay differential equations). Some sections have been rewritten and enlarged. Other changes include new problems, an updated bibliography, and an appendix with pointers to mathematical software. AO

Numerical Analysis, T(15: 2), L. *Practical Numerical Analysis.* Gwynne A. Evans. Wiley, 1995, xiii + 455 pp, $74.95. [ISBN 0-471-95535-3] Numerical techniques for a broad range of problems: solution of linear and nonlinear equations, ordinary and partial differential equations, integral equations; eigenvalue problems; approximation theory; quadrature; optimization. Gives hints and answers for many of the exercises. AO

Numerical Analysis, T*(15–17: 1), C, L. *Fundamentals of Numerical Computing.* L.F. Shampine, R.C. Allen, Jr., S. Pruess. Wiley, 1997, x + 268 pp, $72.95. [ISBN 0-471-16363-5] An introduction to scientific computing requiring only calculus and modest programming experience. Topics reflect problems most common in practice: solving systems of linear and nonlinear equations, interpolation, numerical integration, numerical solution of ODEs. Each chapter includes a case study illustrating the interplay of analysis and computation in the solution of real-world problems. AO

Numerical Analysis, S(15–17), P*, L*. *Afternotes on Numerical Analysis.* G.W. Stewart. SIAM, 1996, x + 200 pp, $29.50 (P). [ISBN 0-89871-362-5] Notes from an undergraduate course taught by a well-known expert; reflects what was actually said in class each day. Topics include nonlinear equations, computer arithmetic, linear equations, polynomial interpolation, numerical integration, and numerical differentiation. Provides valuable insights for students and instructors. AO

Operator Theory, P. *The Asymptotic Behaviour of Semigroups of Linear Operators.* Jan van Neerven. Oper. Theory: Adv. & Applic., V. 88. Birkhäuser Boston, 1996, xii + 237 pp, $122.50. [ISBN 0-8176-5455-0]

Operator Theory, S(18), P, L. *Fundamental Solutions for Differential Operators and Applications.* Prem K. Kythe. Birkhäuser Boston, 1996, xxi + 414 pp, $64.50. [ISBN 0-8176-3869-5] Covers over seventy linear and nonlinear differential operators from mathematical physics, theory of elasticity, fluid dynamics, piezoelectrics, and cosmology. Assumes advanced calculus, ordinary and partial differential equations, complex analysis. KS

Functional Analysis, T(18), S, P, L?. *Banach Spaces for Analysts.* P. Wojtaszczyk. Stud. in Adv. Math., V. 25. Cambridge Univ Pr, 1996, xiii + 382 pp, $34.95 (P); $69.50. [ISBN 0-521-56675-4; 0-521-35618-0] Dense introduction. Emphasizes usefulness for harmonic analysis, complex function theory, orthonormal series, approximation theory, and probability. Lots of exercises, with hints and solutions. Extensive bibliography. A useful, well-written guide. KS

Functional Analysis, P. *Function Spaces, Entropy Numbers and Differential Operators.* David E. Edmunds, Hans Triebel. Tracts in Math., V. 120. Cambridge Univ Pr, 1996, xi + 252 pp, $59.95. [ISBN 0-521-56036-5]

Analysis, P. *Geometry of Harmonic Maps.* Yuanlong Xin. Prog. in Nonlinear Diff. Eqts. & Their Applic., V. 23. Birkhäuser Boston, 1996, x + 241 pp, $79.50. [ISBN 0-8176-3820-2]

Analysis, P. *Ergodic Theory of $\mathbb{Z}^d$-Actions.* Eds: Mark Pollicott, Klaus Schmidt. London Math. Soc. Lect. Note Ser., V. 228. Cambridge Univ Pr, 1996, viii + 484 pp, $44.95 (P). [ISBN 0-521-57688-1] Proceedings of the 1993–94 Warwick Symposium. 8 surveys and 12 research papers.

Analysis, T(16–17: 1), L. *Integral Transformations, Operational Calculus, and Generalized Functions.* R.G. Buschman. Math. & Its Applic., V. 377. Kluwer Academic, 1996, xiii + 231 pp, $117. [ISBN 0-7923-4183-X] Elementary introduction to integral transforms and related topics; requires only calculus and linear algebra background. Topics include Laplace, Fourier, Mellin transforms; Mikusinski operators; generalized functions. Proofs not stressed

but references are given. Computational exercises and more substantial problems. Unfortunate misspellings in "Contents." JS

Analysis, T(16–17: 1), S, L. *Transform Methods in Applied Mathematics: An Introduction.* Peter Lancaster, Kęstutis Šalkauskas. Canadian Math. Soc. Ser. of Mono. & Adv. Texts. Wiley, 1996, x + 332 pp, $59.95. [ISBN 0-471-00810-9] Introduction to significant modern applications of transform theory. Required background is limited to calculus and linear algebra; includes two chapters on complex variables. Topics include the Laplace transform, Fourier methods, Z-transforms, discrete and continuous filters, distributions, and wavelets. JS

Analysis, P. *Regularization of Inverse Problems.* Heinz W. Engl, Martin Hanke, Andreas Neubauer. Math. & Its Applic., V. 375. Kluwer Academic, 1996, viii + 321 pp, $160. [ISBN 0-7923-4157-0]

Analysis, P. *Algebraic K-Theory.* Eds: Grzegorz Banaszak, Wojciech Gajda, Piotr Krasoń. Contemp. Math., V. 199. AMS, 1996, xviii + 210 pp, $49 (P). [ISBN 0-8218-0511-8] Proceedings of a 1995 conference at the Adam Mickiewicz University in Poznań, Poland.

Algebraic Geometry, T(18: 1), P. *Prolegomena to a Middlebrow Arithmetic of Curves of Genus 2.* J.W.S. Cassels, E.V. Flynn. London Math. Soc. Lect. Note Ser., V. 230. Cambridge Univ Pr, 1996, xiv + 219 pp, $37.95 (P). [ISBN 0-521-48370-0] Curves of genus 2 can be represented as y^2 = a polynomial of degree 6. Most of what is known about them is general theory that is not very helpful when faced with a specific curve. The authors begin filling in this gap by constructing Mordell–Weil groups and finding rational points. DB

Algebraic Geometry, P. *Complex Algebraic Surfaces, Second Edition.* Arnaud Beauville. London Math. Soc. Stud. Texts, V. 34. Cambridge Univ Pr, 1996, ix + 132 pp, $22.95 (P); $59.95. [ISBN 0-521-49842-2; 0-521-49510-5] Reprint of the *First Edition* (TR, November 1984), with a new appendix on complex surfaces. PG

Algebraic Geometry, P. *Spinning Tops: A Course on Integrable Systems.* Michèle Audin. Stud. in Adv. Math., V. 51. Cambridge U Pr, 1996, viii + 139 pp, $34.95. [ISBN 0-521-56129-9]

Differential Geometry, T?(18: 1), P. *Singular Semi-Riemannian Geometry.* Demir N. Kupeli. Math. & Its Applic., V. 366. Kluwer Academic, 1996, x + 177 pp, $99. [ISBN 0-7923-3996-7] Relatively self-contained study of smooth manifolds furnished with degenerate (singular) metric tensors, leading from an intrinsic view to an extrinsic one. Topics include reasonable preliminaries, singular (and singular quaternionic) Kahler manifolds. RM

Geometry, P. *Geometric Applications of Fourier Series and Spherical Harmonics.* H. Groemer. Ency. of Math. & Its Applic., V. 61. Cambridge Univ Pr, 1996, xi + 329 pp, $59.95. [ISBN 0-521-47318-7] Proves results from geometry (mostly from convex set theory) using Fourier series and spherical harmonics. Includes background material on convex sets, analysis; develops necessary results on Fourier series, spherical harmonics. LC

Algebraic Topology, P. *Ends of Complexes.* Bruce Hughes, Andrew Ranicki. Tracts in Math., V. 123. Cambridge Univ Pr, 1996, xxv + 353 pp, $64.95. [ISBN 0-521-57625-3]

Topology, T(15), L. *Knot Theory and Its Applications.* Kunio Murasugi. Transl: Bohdan Kurpita. Birkhäuser Boston, 1996, 341 pp, $69.50. [ISBN 0-8176-3817-2] Introductory treatment. Little background assumed, so no mention of algebraic topology. Plenty of pictures and exercises. Several sections show applications to science. JD

Topology, T(17–18: 1). *Topology Via Logic.* Steven Vickers. Tracts in Theoret. Comp. Sci., V. 5. Cambridge Univ Pr, 1996, 200 pp, $24.95 (P). [ISBN 0-521-57651-2] Paperback republication of 1989 edition (TR, March 1990). Written for computer scientists.

Operations Research, T(17: 2), P, L*. *Monte Carlo: Concepts, Algorithms, and Applications.* George S. Fishman. Ser. in Oper. Res. Springer-Verlag, 1996, xxv + 698 pp, $69. [ISBN 0-387-94527-X] A comprehensive introduction with chapters on estimating volume and count, generating samples, increasing efficiency, random tours, designing and analyzing sample paths, and generating pseudorandom numbers. AO

Operations Research, T?(16: 1), C. *Introduction to Practical Linear Programming.* David J. Pannell. Wiley, 1997, xiii + 333 pp, $59.95, with disk. [ISBN 0-471-51789-5] Discusses model construction, interpretation of software output, sensitivity analysis, etc. Nonmathematical—the simplex algorithm is not presented. AO

Optimization, T(17: 1), C. *A First Course in Optimization Theory.* Rangarajan K. Sundaram. Cambridge Univ Pr, 1996, xvii + 357 pp, $74.95; $27.95 (P). [ISBN 0-521-49719-1; 0-521-49770-1] Blends theory and economic

applications. Major topics: existence of solutions in $\mathbb{R}^n$ (e.g., theorems of Weierstrass and Lagrange, Karush–Kuhn–Tucker conditions); parametric variation of solutions; finite- and infinite-horizon dynamic programming. AO

Optimization, T(16–17: 1), C. *Linear and Integer Programming: Theory and Practice.* Gerard Sierksma. Pure & Appl. Math., V. 198. Marcel Dekker, 1996, xiv + 673 pp, $175, with disk. [ISBN 0-8247-9695-0] Covers the simplex method, duality and sensitivity analysis, basic solution techniques for integer programming problems, and the interior path version of Karmarkar's interior point method. Includes a chapter on model building with case studies; proofs of major results. Exercise solutions are in an appendix. AO

Optimal Control, T(18: 1), P. *Operator Approach to Linear Control Systems.* A. Cheremensky, V. Fomin. Math. & Its Applic., V. 345. Kluwer Academic, 1996, xvi + 396 pp, $198. [ISBN 0-7923-3765-4] Develops operator theoretic approach to LQP (optimization problems for linear control with quadratic performance index). Extends classical theory to cases with distributed parameter plant described by functional (time delay) or partial differential equations. The linear control plant is treated as an unbounded linear operator in Hilbert space with time structure. RM

Probability, T(18: 1), P. *Lévy Processes.* Jean Bertoin. Tracts in Math., V. 121. Cambridge Univ Pr, 1996, x + 265 pp, $54.95. [ISBN 0-521-56243-0] Concise overview of stochastic processes with independent and stationary time increments (random walks in continuous time) within Euclidean framework, using analytic tools (Fourier, Laplace transforms). These include Poisson processes, Brownian motion, stable processes, and serve as prototypes of Markov processes. RM

Stochastic Processes, P. *Continuum Percolation.* Ronald Meester, Rahul Roy. Tracts in Math., V. 119. Cambridge Univ Pr, 1996, x + 238 pp, $49.95. [ISBN 0-521-47504-X]

Mathematical Statistics, T(17: 1). *A Course in Large Sample Theory.* Thomas S. Ferguson. Texts in Stat. Sci. Chapman & Hall, 1996, ix + 245 pp, $34.95 (P). [ISBN 0-412-04371-8] Covers basic probability, laws of large numbers, central limit theorems, Slutsky theorems, asymptotic theory and distributions of quantiles, maximum likelihood estimates, likelihood ratio statistics, and posterior distributions. Contains many examples and exercises with detailed solutions. RS

Statistical Methods, T(18), P. *Smoothing Methods in Statistics.* Jeffrey S. Simonoff. Ser. in Stat. Springer-Verlag, 1996, xii + 338 pp, $54.95. [ISBN 0-387-94716-7] Discusses a variety of methods for nonparametric density, distribution, and regression function estimation. Includes older as well as recently developed techniques. Highlights useful and promising methods for scientists who analyze data from a practical standpoint. Most topics motivated by data examples. Computational issues section lists sources of code for methods. RS

Statistical Methods, P. *Advances in Biometry: 50 Years of the International Biometric Society.* Eds: Peter Armitage, Herbert A. David. Ser. in Prob. & Stat. Wiley, 1996, xii + 473 pp, $59.95. [ISBN 0-471-16018-0] 21 articles covering a wide range of topics including mathematical and statistical techniques, methodology, and applications in various fields of biology and medicine. RS

Statistical Methods, T(17-18: 1, 2), P. *Methods and Applications of Linear Models: Regression and the Analysis of Variance.* Ronald R. Hocking. Ser. in Prob. & Stat. Wiley, 1996, xxii + 731 pp, $69.95. [ISBN 0-471-59282-X] Introduces linear model theory then discusses linear regression models and analysis of variance models. Graphical examples used throughout to illustrate and simplify theory. Presents classical methods as well as new and developing techniques (e.g., Hocking's AVE method for mixed effects models). RS

Statistical Methods, P. *Adaptive Sampling.* Steven K. Thompson, George F. Seber. Ser. in Prob. & Stat. Wiley, 1996, xi + 265 pp, $54.95. [ISBN 0-471-55871-0] Assumes some knowledge of sample survey theory. Discusses a wide range of methods including recent developments. Worked examples are given throughout to illustrate theory and methods. RS

Mathematical Computing, T*(15–16), S. *Maple: A Comprehensive Introduction.* Roy Nicolaides, Noel Walkington. Cambridge Univ Pr, 1996, xix + 466 pp, $39.95. [ISBN 0-521-56230-9] An introduction to Maple at a somewhat deeper level than the average "Intro to Maple" text. Emphasizes why and how Maple works, not just what it does. Aims to provide the reader with a solid foundation in Maple as a computer algebra system and a programming language. Appears to succeed admirably. Covers Maple V Release 4. MPR

Mathematical Computing, P, C. *Computation of Special Functions.* Shanjie Zhang, Jianming Jin. Wiley, 1996, xxvi + 717 pp, $89.95, with disk. [ISBN 0-471-11963-6] Software

for evaluating special functions (or some property of them, such as their derivatives, zeros, coefficients of certain expansions, etc.) that arise in engineering or the sciences. LC

Mathematical Computing, S(15–17), P*, L*. *Introduction to Maple, Second Edition.* André Heck. Springer-Verlag, 1996, xx + 699 pp, $39.95. [ISBN 0-387-94535-0] Revised and updated to reflect new mathematical features and improvements in Maple V Release 4. (*First Edition*, TR, March 1994.) AO

Computer Science, T(16–17: 1, 2), L. *Foundations for Programming Languages.* John C. Mitchell. Found. of Comp. Ser. MIT Pr, 1996, xix + 846 pp, $60. [ISBN 0-262-13321-0] Modern, rich, and encyclopedic treatment of the logico-mathematical foundations of programming languages. Stresses analysis of syntactic, operational, and semantic properties via a series of typed lambda calculi. Useful as an introduction to the foundations of languages, semantics, or type theory; deep but accessible to good undergraduates. RM

Applications (Economics), P. *Mathematical Models in Finance.* Eds: S.D. Howison, F.P. Kelly, P. Wilmott. Chapman & Hall, 1995, viii + 152 pp, $60.50. [ISBN 0-412-63070-2] 13 papers from a Royal Society of London discussion meeting. Subjects range from classical theory to recent research results.

Applications (Economics), P, C. *Computational Economics and Finance: Modeling and Analysis with Mathematica.* Ed: Hal R. Varian. Springer-Verlag, 1996, xiv + 468 pp, $54.95, with disk. [ISBN 0-387-94518-0] 16 articles illustrate use of *Mathematica* in economics, finance, and statistics. Diskette contains related *Mathematica* notebooks and packages.

Applications (Economics), T(14: 1), S. *Mathematics for Economics and Finance: Methods and Modelling.* Martin Anthony, Norman Biggs. Cambridge Univ Pr, 1996, 394 pp, $24.95 (P); $75. [ISBN 0-521-55913-8; 0-521-55113-7] An introduction to calculus (single- and multivariable) and linear algebra in the context of economics. Each chapter organized as narrative exposition, worked examples, summary (list of main topics, key terms, notations, and formulae), and exercises. AO

Applications (Economics), P. *An Introduction to Bayesian Inference in Econometrics.* Arnold Zellner. Classics Lib. Ed. Wiley, 1996, xv + 431 pp, $39.95 (P). [ISBN 0-471-16937-4] Paperback republication of 1971 edition (TR, April 1973).

Applications (Economics), T(17: 1), P. *Introduction to Stochastic Calculus Applied to Finance.* Damien Lamberton, Bernard Lapeyre. Transl: Nicolas Rabeau, François Mantion. Chapman & Hall, 1996, xi + 185 pp, $45.95. [ISBN 0-412-71800-6] Applications of probabilistic techniques to financial modeling. Presents discrete-time models, the Black–Scholes model, interest rate models, and simulation techniques. Assumes background in measure-theoretic probability. AO

Applications (Mechanics), P. *Lecture Notes in Control and Information Sciences–220: Nonsmooth Impact Mechanics: Models, Dynamics and Control.* Bernard Brogliato. Springer-Verlag, 1996, xv + 400 pp, $76 (P). [ISBN 3-540-76079-2]

Applications (Quantum Theory), P. *The Infamous Boundary: Seven Decades of Heresy in Quantum Physics.* David Wick. Springer-Verlag, 1995, xvii + 310 pp, $19 (P). [ISBN 0-387-94726-4] Paperback republication of 1995 Birkhäuser Boston edition (TR, March 1996).

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–217: Robust Control via Variable Structure and Lyapunov Techniques.* Eds: Franco Garofalo, Luigi Glielmo. Springer-Verlag, 1996, xxi + 307 pp, $63 (P). [ISBN 3-540-76067-9] Papers from a 1994 IEEE Workshop in Benevento, Italy.

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–221: Control of Nonlinear Multibody Flexible Space Structures.* Atul Kelkar, Suresh Joshi. Springer-Verlag, 1996, xiv + 141 pp, $43 (P). [ISBN 3-540-76093-8]

Applications, P. *Computational Psycholinguistics: AI and Connectionist Models of Human Language Processing.* Eds: Ton Dijkstra, Koenraad de Smedt. Taylor & Francis, 1996, xv + 437 pp, $29.95 (P). [ISBN 0-7484-0466-X]

Applications, T(13: 1). *Early Astronomy.* Hugh Thurston. Springer-Verlag, 1994, x + 268 pp, $29.95 (P). [ISBN 0-387-94107-X] A very enjoyable account of early developments in astronomy. PF

Reviewers

DB: David Bressoud, Macalester; LC: Laura Chihara, St. Olaf; JD: Jill Dietz, St. Olaf; PF: Paul Froeschl, Macalester; PG: Philip Gloor, St. Olaf; TH: Tom Halverson, Macalester; DK: Danny Kaplan; Macalester; RM: Richard Molnar, Macalester; AO: Arnold Ostebee, St. Olaf; MPR: Matthew P. Richey, St. Olaf; KS: Karen Saxe, Macalester; JS: John Schue, Macalester; RS: Richard Single, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf.

THE AUTHORS

A. J. BERRICK was educated at the University of Sydney and at Oxford University where he became a fellow of St. John's College. After lecturing at Imperial College, London, he joined the National University of Singapore's Mathematics Department, where he is now a professor. His research publications have mainly been in algebraic topology, cohomology of groups, and algebraic K-theory, the subject of his Pitman book.

M. E. KEATING obtained his B.Sc. at University College, London, and his Ph.D. at King's College, London. He then joined the Mathematics Department at Imperial College, London, where he has been a permanent fixture apart from sabbaticals at McGill University, Montreal, and the National University of Singapore. His research interests center on the relationships between number theory, noncommutative ring theory, and K-theory. While Berrick and Keating were together in Singapore, they started a long-term collaboration to produce graduate level texts on K-theory and its prerequisites, which should soon bear fruit in the Cambridge University Press.

MAARTEN BOERLIJST, born 1963, received his Ph.D. in 1994 with Pauline Hogeweg in Utrecht (the Netherlands). His research interests include prebiotic evolution, spatial models in ecology and evolution, theoretical immunology, virus dynamics, and evolutionary game theory. He is now a "European Commission Human Capital and Mobility Fellow" or "Flying Dutchman" with the Mathematical Biology group in Oxford. He enjoys outdoor sports, including sailing, horseback riding, and travelling.

MARTIN NOWAK was born in Vienna in 1965, met Karl Sigmund in 1987, and became immediately engaged in an (almost) endlessly repeated Prisoner's Dilemma. Nowak eventually escaped to Oxford, where he is Head of the Mathematical Biology group in the Zoology Department. His research interests range from mathematical models of evolutionary dynamics to virus infections and the immune system.

KARL SIGMUND, born 1945, is professor of mathematics at the University of Vienna. His interests include evolutionary game theory, dynamical systems, and (more recently) the history of the Vienna Circle. His main attempt at popular science is the book *Games of Life*.

JIM SAUERBERG grew up in northern Wisconsin (Twins), and obtained a B.S. from the University of Wisconsin at Madison (Brewers) and a Ph.D. from Brown University (Red Sox). He then went to Union College in Schenectady, New York (None nearby–argh!), and has recently moved to St. Mary's College in the Bay Area (Giants and A's). He manages to do some number theory between innings.

LINGHSUEH SHU received degrees in mathematics from National Central University in Taiwan and from Brown University, and has since taught at the Ohio State University and the University of Vermont. Although her expertise is in number theory, and she enjoys all types of mathematics, she is beginning to suspect that her true love is pottery.

DAN VELLEMAN received his bachelor's degree from Dartmouth College in 1976 and his doctorate from the University of Wisconsin in 1980. He taught at the University of Texas and the University of Toronto before joining the faculty of Amherst College in 1983. He is interested in logic, the philosophy of mathematics, and the foundations of quantum mechanics. When he isn't doing mathematics he enjoys bicycling and singing.

ALFRED M. BRUCKSTEIN was born in Transylvania, then moved to Israel, and is now at Bell Laboratories for an extended sabbatical from the Technion in Haifa, where he is Professor of Computer Science. He became interested in modeling ants and their trails while reading "Surely You are Joking, Mr. Feynman." He enjoys designing logos and, most of all, playing with his son, Ariel.

COLIN L. MALLOWS has recently left Bell Labs (which is now part of Lucent Technologies) and joined the new AT & T Laboratory, where he plans to continue to try to make statistics seem less like sorcery.

ISRAEL A. WAGNER is a research staff member at IBM Haifa Research Laboratory, where he creates and chases bugs in VLSI circuits for high-end processors. He is also working on a Ph.D. thesis, trying to show that many small-scale agents can cooperate to achieve large-scale targets.

RICHARD STANLEY received his Ph.D. from Harvard University under the direction of Gian-Carlo Rota (at M.I.T.). He was a Moore Instructor at M.I.T. in 1970–1971 and a Miller Research Fellow at Berkeley in 1971–1973. In 1973 he returned to M.I.T. where he is now Professor of Applied Mathematics. Stanley's main research interest is combinatorics, especially its connections with algebra, geometry, and topology. He has long been interested in classical history and is pleased to have the opportunity to combine his interest in this area with that of mathematics.

GERT ALMKVIST received his Ph.D. from UC Berkeley in 1966. He has been at University of Lund since 1967 with visits to Berkeley and Urbana. His main interests are in algebraic K-theory, invariant theory, and analytic number theory. He is the founder of the Institute of Algebraic Meditation. He enjoys table tennis (once commutative algebraic world champion), tennis, and hiking.

MARTIN L. JONES is Associate Professor of Mathematics at the University of Charleston, South Carolina where he has been since receiving his Ph.D. in mathematics from the Georgia Institute of Technology in 1989. He has recently returned from spending a year as a Fulbright Scholar at the University of the Andes in Venezuela. His research interests are in sequential decision theory, most notably optimal stopping problems and bandit processes.

HASSAN SEDAGHAT received his Ph.D. from the George Washington University, where he worked on semigroup compactifications. Nowadays, he is studying the global behavior of trajectories of nonlinear scalar and vector difference equations. He is also interested in the applications of such equations in economics, particularly, in the dynamical modelling of consumer demand without utility functions. His nonmathematical interests include reading classical fiction and watching adventure and science fiction movies.

JET WIMP, a mathematician and a poet, teaches at Drexel University in Philadelphia.

FRANK MORGAN works in minimal surfaces. He has a weekly live call-in Math Chat on local cable TV, featured in Ivars Peterson's column MathLand at the MAA web site at http://www.maa.org/(search Frank and Morgan:BY), and a biweekly Math Chat column in the *Christian Science Monitor*, sometimes available via the web page http://www.csmonitor.com/. His books include *Geometric Measure Theory: a Beginner's Guide*, *Riemannian Geometry: a Beginner's Guide*, and *Calculus Lite*.

EDITOR'S ENDNOTES

Michael Hardy sent the following comment after reading David Poole's article The Stochastic Group [**102** (1995) 798–801]:

> David Poole showed that the set of $n \times n$ nonsingular matrices in which the sum of the entries in each column is 1 is a group, and is isomorphic to the affine group that is often represented by the group of matrices of the form $\begin{pmatrix} A & \mathbf{b}^T \\ 0 & 1 \end{pmatrix}$. Poole's group of "stochastic matrices" deserves to be called the "natural representation of the affine group." An "affine combination" of n points $\mathbf{p}_1, \ldots, \mathbf{p}_n$ in an affine space is a point $c_1\mathbf{p}_1 + \cdots + c_n\mathbf{p}_n$, where the scalars $c, \ldots, c_n$, satisfy $c_1 + \cdots + c_n = 1$. Consider n affinely independent points $\mathbf{p}_1, \ldots, \mathbf{p}_n$ in an $(n-1)$-dimensional affine space. Any affine automorphism γ is determined by $\gamma(\mathbf{p}_1), \ldots, \gamma(\mathbf{p}_n)$, and every point in the affine space can be written uniquely as an affine combination of $\mathbf{p}_1, \ldots, \mathbf{p}_n$. Thus $\gamma(\mathbf{p}_j) = c_{1j}\mathbf{p}_1 + \cdots + c_{nj}\mathbf{p}_n$ for $j = 1, \ldots, n$. The matrix C whose i, j entry is c_{ij} represents this affine transformation. Identify any point $\mathbf{d}$ in the affine space with the column vector of its affine coefficients, writing $\mathbf{d} = d_1\mathbf{p}_1 + \cdots + d_n\mathbf{p}_n = (d_1, \ldots, d_n)^T$. The (column vector of affine coefficients of the) image of $\mathbf{d}$ under γ is $\gamma(\mathbf{d}) = C\mathbf{d}$. It follows that matrix multiplication corresponds to composition of affine transformations, so this really is a group representation. All of this works just as well regardless of the field of scalars. There is also a geometric, rather than algebraic, way to look at the argument showing that the two groups of matrices are conjugate to each other as subgroups of $GL(n)$: Think about which $(n-1)$-dimensional affine subspaces are invariant

In response, Poole said that Hardy's observations "nicely complement the main result of my article ... I completely agree ... that the representation ... is the 'natural' one for the affine group."

Two printer's errors mar the displayed inequality in part (*ii*) of the Theorem at the top of page 804 in J. W. Sander's article A Story of Binomial Coefficients and Primes [**102** (1995) 802–807]. The correct inequality is

$$E_{a,p}(p^k) \leq \left(\frac{p+1}{2}\right)^k \sum_{t=0}^{a-1} \binom{k}{t}.$$

Paul Smith, writing about Arthur T. White's article on Fabian Stedman [**103** (1996) 771–778], says

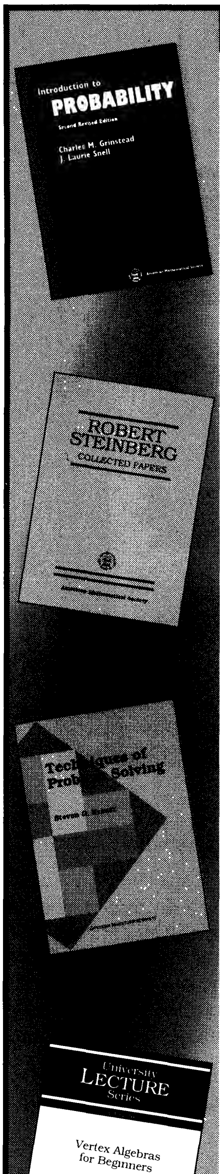
> White makes a convincing case for an implicit mastery of group theory by the composer of Stedman Doubles—a change (composition) for bell ringers. This is not the first time that insight was offered in the MONTHLY, however. T. J. Fletcher, in Campanological Groups [**63** (1956) 619–626] observed: "One of the most pleasing methods from a mathematical point of view is *Stedman's Doubles*. This method was invented round about 1640, and it displays a striking knowledge of decomposition into cosets very nearly one hundred years before Lagrange was born."

Alan Horwitz wrote

> The article A Hundred Years of Prime Numbers [**103** (1996) 729–741] looks very interesting, but shouldn't the MONTHLY have waited one year and celebrated 101 years of prime numbers?

The editor regrets the misspelling of the name of the lyricist for the delightful "It's De-Lovely" in last November's issue [**103** (1996) 770]. He is Ronald E. Prather, Trinity University.

Roger A. Horn, *Editor*



American Mathematical Society

Recently Published by the AMS

Introduction to Probability

Second Revised Edition

Charles M. Grinstead, *Swarthmore College, PA,* and **J. Laurie Snell,** *Dartmouth College, Hanover, NH*

This text is designed for an introductory probability course at the university level for sophomores, juniors, and seniors in mathematics, the physical and social sciences, engineering, and computer science. It presents a thorough treatment of probability ideas and techniques necessary for a firm understanding of the subject.

The text is also recommended for use in discrete probability courses. The material is organized so that the discrete and continuous probability discussions are presented in a separate, but parallel, manner. This organization doesn't emphasize an overly rigorous or formal view of probabililty and therefore offers some strong pedagogical value. Hence, the discrete discussions can sometimes serve to motivate the more abstract continuous probability discussions.

Features:

- Key ideas are developed in a somewhat leisurely style, providing a variety of interesting applications to probability and showing some nonintuitive ideas.
- Over 600 exercises provide the opportunity for practicing skills and developing a sound understanding of ideas.
- Text includes many computer programs that illustrate the algorithms or the methods of computation for important problems.

1997; 484 pages; Hardcover; ISBN 0-8218-0749-8; List $49; All AMS members $39; Order code IPROBMM74

Robert Steinberg Collected Papers

Robert Steinberg, *University of California, Los Angeles*

This volume contains all of Steinberg's published papers on group theory, including those on "special representations" (now called Steinberg representations), tensor products of representations, finite reflection groups, regular elements of algebraic groups, Galois cohomology, universal extensions, etc. At the end of the book, there is a section called "Comments on the Papers". The comments by Steinberg contain minor corrections and clarifications and explain how ideas and results have evolved and been used since they first appeared.

Collected Works, Volume 7; 1997; 598 pages; Hardcover; ISBN 0-8218-0576-2; List $79; Individual member $47; Order code CWORKS/7MM74

Techniques of Problem Solving

Steven G. Krantz, *Washington University, St. Louis, MO*

... the subject of problem solving ... is more than just a disconnected list of brain teasers and recreations. It is a way of life. Scientists of every stripe—chemists, physicists, psychologists, social engineers, and many others—ply their trade by considering a set of data, deciding what techniques are relevant to these data, and then solving a problem. *It is this view of problem solving that will be promulgated in the present book.*

—from the Preface

The purpose of this book is to teach the basic principles of problem solving, including both mathematical and nonmathematical problems. This book will help students to ...

- translate verbal discussions into analytical data.
- learn problem-solving methods for attacking collections of analytical questions or data.
- build a personal arsenal of internalized problem-solving techniques and solutions.
- become "armed problem solvers", ready to do battle with a variety of puzzles in different areas of life.

Taking a direct and practical approach to the subject matter, Krantz's book stands apart from others like it in that it incorporates exercises throughout the text. After many solved problems are given, a "Challenge Problem" is presented. Additional problems are included for readers to tackle at the end of each chapter. There are more than 350 problems in all. A *Solutions Manual* to most end-of-chapter exercises is available.

1997; 465 pages; Softcover; ISBN 0-8218-0619-X; List $29; All AMS members $23; Order code TPSMM74

Vertex Algebras for Beginners

Victor Kac, *Massachusetts Institute of Technology, Cambridge*

This book is an introduction to algebraic aspects of conformal field theory, which in the past decade revealed a variety of unusual mathematical notions. Vertex algebra theory provides an effective tool to study them in a unified way.

Here, a mathematician will encounter new algebraic structures that originated from Einstein's special relativity postulate and Heisenberg's uncertainty principle. A physicist will find familiar notions presented in a more rigorous and systematic way, which may lead to a better understanding of foundations of quantum physics.

University Lecture Series, Volume 10; 1997; 141 pages; Softcover; ISBN 0-8218-0643-2; List $25; All AMS members $20; Order code ULECT/10MM74

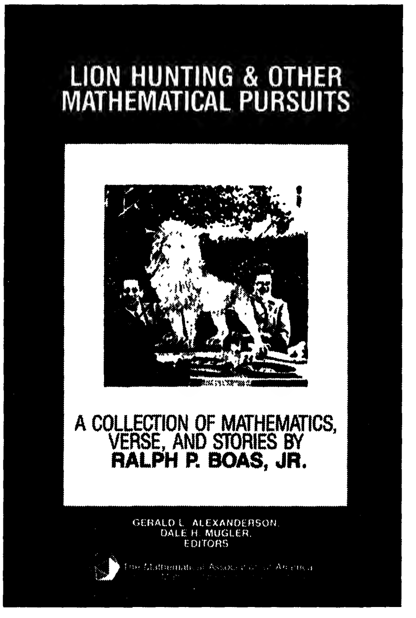
All prices subject to change. Charges for delivery are $3.00 per order. For air delivery outside of the continental U. S., please include $6.50 per item. *Prepayment required.* Order from: **American Mathematical Society**, P. O. Box 5904, Boston, MA 02206-5904. For credit card orders, fax (401) 331-3842 or call toll free 800-321-4AMS (4267) in the U. S. and Canada, (401) 455-4000 worldwide. Or place your order through the AMS bookstore at http://www.ams.org/bookstore/. Residents of Canada, please include 7% GST.

Lion Hunting

and Other Mathematical Pursuits

A Collection of Mathematics, Verse, and Stories by Ralph P. Boas, Jr.

Gerald L. Alexanderson and Dale H. Mugler, Editors



I highly recommend Lion Hunting and Other Mathematical Pursuits to high school mathematics clubs, mathematics teachers of all levels, and anyone interested in mathematics. Perhaps the most important features of this book is how it subtly makes the reader aware of the nature of mathematics.

– The Mathematics Teacher

As a young man at the Institute for Advanced Study in Princeton, Ralph Philip Boas, Jr., together with a group of other mathematicians, published a light-hearted article on the "mathematics of lion hunting" under a pseudonym (1938). This sparked a sequence of articles on the topic, several of which are drawn together in this book.

Lion Hunting includes an assortment of articles that show the many facets of this remarkable mathematician, editor, writer, and teacher. Along with a variety of his lighter mathematical papers, the collection includes Boas' verse and short stories, many of which are appearing for the first time. Anecdotes and recollections of his numerous experiences and of his work and meetings with many distinguished mathematicians and scientists of his day are also included as well as photographs taken by Boas of Hardy, Littlewood, Besicovitch, Weil, and others.

The mathematical articles in this collection cover a range of topics. They include articles on infinite series, the mean value theorem, indeterminate forms, complex variables, inverse functions, extremal problems for polynomials and more. A special section of this book is devoted to articles about the teaching of mathematics, with titles such as "Calculus as an experimental science" and "Can we make mathematics intelligible?"

Boas's wit and playful humor are reflected in the verses included in this collection. The verses reflect the phases of his career as author, editor, teacher, department chair, and lover of literature. A section of the book describes the feud that Boas supposedly had with Bourbaki. Also included are many amusing anecdotes about famous mathematicians.

320 pp., Paperbound, 1995, ISBN 0-88385-323-X
List: $39.95 MAA Member: $28.95
Catalog Code: DOL-15

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
P.O. Box 91112, Washington, DC 20090-1112
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

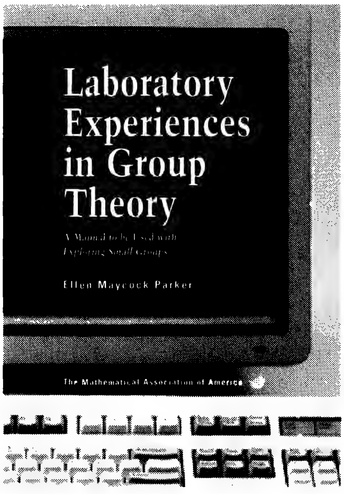
QTY.	CATALOG CODE	PRICE	AMOUNT
______	DOL-15	______	______
______	______	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

THE MATHEMATICAL ASSOCIATION OF AMERICA



Laboratory Experiences in Group Theory

A Manual to be Used with *Exploring Small Groups*

Ellen Maycock Parker

Series: Classroom Resource Materials

A lab manual with software for introductory courses in group theory or abstract algebra

Laboratory Experiences in Group Theory is a workbook of 15 laboratories designed to be used with the software *Exploring Small Groups* as a supplement to the regular textbook in an introductory course in group theory or abstract algebra. Written in a step-by-step manner, the laboratories encourage students to discover the basic concepts of group theory and to make conjectures from examples that are easily generated by the software. The labs can be assigned as homework or can be used in a structured laboratory setting. Since the software is user-friendly and the laboratories are complete, students and faculty should have no difficulty in using the labs without training.

Most students find that the laboratories provide an enjoyable alternative to the "theorem-proof-example" format of a standard abstract algebra course. At the end of the semester, one student wrote in his evaluation of the course:

I am truly grateful for the laboratory component...Work on the computer helped to make the abstract theory more concrete... One of the best things about the labs was that we formed our own conjectures about the patterns we saw...I believe that the progression of (1) lab, (2) conjecture, (3) class discussion, and (4) proof was highly beneficial in gaining understanding of the abstract material of the course.

Table of Contents: 1. Groups and Geometry; 2. Cayley Tables; 3. Cyclic Groups and Cyclic Subgroups; 4. Subgroups and Subgroup Lattices; 5. The Center and Commutator Subgroups; 6. Quotient Groups; 7. Direct Products; 8. The Unitary Groups; 9. Composition Series; 10. Introduction to Endomorphisms; 11. The Inner Automorphisms of a Group; 12. The Kernel of an Endomorphism; 13. The Class Equation; 14. Conjugate Subgroups; 15. The Sylow Theorems; Appendix A. Table Generation Menu of *Exploring Small Groups (ESG)*; Appendix B. Sample Library of *ESG*; Appendix C. Group Library of *ESG*; Appendix D. Group Properties Menu

Exploring Small Groups, the software packaged with this lab manual, is on a 3½" DD PC compatible disk. This is a DOS program that can be run in Windows. The software was developed by Ladnor Geissinger, University of North Carolina at Chapel Hill.

112 pp., Paperbound, 1996
ISBN 0-88385-705-7
List: $22.00 MAA Member: $16.00
Catalog Code: LABEJR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
PO Box 91112, Washington, DC 20090-1112
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
____	LABE/JR	____	____
____	____	____	____
		TOTAL	____

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

THE AMERICAN MATHEMATICAL MONTHLY

Volume 104, Number 5 May 1997

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and Notes should be sent to the Editor:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send your email address and 3 copies of the complete manuscript (including all figures with captions and lettering), typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper.

Letters to the Editor on any topic are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

See the MONTHLY section of MAA Online for current information such as contents of issues, descriptive summaries of forthcoming articles, tips for authors, and preparation of manuscripts in TEX:

http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all material, typewritten on only one side of the paper.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Mr. JOE WATSON, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Was Newton's Calculus a Dead End? The Continental Influence of Maclaurin's Treatise of Fluxions

Judith V. Grabiner

1. INTRODUCTION. Eighteenth-century Scotland was an internationally-recognized center of knowledge, "a modern Athens in the eyes of an enlightened world." [**74**, p. 40] [**81**] The importance of science, of the city of Edinburgh, and of the universities in the Scottish Enlightenment has often been recounted. Yet a key figure, Colin Maclaurin (1698–1746), has not been highly rated. It has become a commonplace not only that Maclaurin did little to advance the calculus, but that he did much to retard mathematics in Britain—although he had (fortunately) no influence on the Continent. Standard histories have viewed Maclaurin's major mathematical work, the two-volume *Treatise of Fluxions* of 1742, as an unread monument to ancient geometry and as a roadblock to progress in analysis. Nowadays, few people read the *Treatise of Fluxions*. Much of the literature on the history of the calculus in the eighteenth and nineteenth centuries implies that few people read it in 1742 either, and that it marked the end—the dead end—of the Newtonian tradition in calculus. [**9**, p. 235], [**49**, p. 429], [**10**, p. 187], [**11**, pp. 228–9], [**43**, pp. 246–7], [**42**, p. 78], [**64**, p. 144]

But can this all be true? Could nobody on the Continent have cared to read the major work of the leading mathematician in eighteenth-century Scotland? Or, if the work was read, could it truly have been "of little use for the researcher" [**42**, p. 78] and have had "no influence on the development of mathematics"? [**64**, p. 144]

We will show that Maclaurin's *Treatise of Fluxions* did develop important ideas and techniques and that it did influence the mainstream of mathematics. The Newtonian tradition in calculus did not come to an end in Maclaurin's Britain. Instead, Maclaurin's *Treatise* served to transmit Newtonian ideas in calculus, improved and expanded, to the Continent. We will look at what these ideas were, what Maclaurin did with them, and what happened to this work afterwards. Then, we will ask what by then should be an interesting question: why has Maclaurin's role been so consistently underrated? These questions will involve general matters of history and historical writing as well as the development of mathematics, and will illustrate the inseparability of the external and internal approaches in understanding the history of science.

2. THE STANDARD PICTURE. Let us begin by reviewing the standard story about Maclaurin and his *Treatise of Fluxions*. The calculus was invented independently by Newton and Leibniz in the late seventeenth century. Newton and Leibniz developed general concepts—differential and integral for Leibniz, fluxion and fluent for Newton—and devised notation that made it easy to use these concepts. Also, they found and proved what we now call the Fundamental Theorem of Calculus, which related the two main concepts. Last but not least, they successfully applied their ideas and techniques to a wide range of important problems.

[**9**, p. 299] It was not until the nineteenth century, however, that the basic concepts were given a rigorous foundation.

In 1734 George Berkeley, later Bishop of Cloyne, attacked the logical validity of the calculus as part of his general assault on Newtonianism. [**12**, p. 213] Berkeley's criticisms of the rigor of the calculus were witty, unkind, and—with respect to the mathematical practices he was criticizing—essentially correct. [**6**, v. 4, pp. 65–102] [**38**, pp. 33–34] [**82**, pp. 332–338] Maclaurin's *Treatise* was supposedly intended to refute Berkeley by showing that Newton's calculus was rigorous because it could be reduced to the methods of Greek geometry. [**10**, pp. 181–2, 187] [**9**, pp. 233, 235] Maclaurin himself said in this preface that he began the book to answer Berkeley's attack, [**63**, p. i] and also to rebut Berkeley's accusation that mathematicians were hostile to religion. [**78**, p. 50]

The majority of Maclaurin's treatise is contained in its first Book, which is called "The Elements of the Method of Fluxions, Demonstrated after the Manner of the Ancient Geometricians." That title certainly sounds as though it looks backward to the Greeks, not forward to modern analysis. And the text is full of words—lots of words. So much time is spent on preliminaries that it is not until page 162 that he can show that the fluxion of **ay** is **a** times the fluxion of **y**. Florian Cajori, whose writings have helped spread the standard story, compared Maclaurin to the German poet Klopstock who, Cajori said, was praised by all, read by none. [**10**, p. 188] While British mathematicians, bogged down with geometric baggage, studied and revered the work and notation of Newton and argued with Berkeley over foundations, Continental mathematicians went onward and upward analytically with the calculus of Leibniz. The powerful analytic results and techniques in eighteenth-century Continental mathematics were all that mathematicians like Cauchy, Riemann, and Weierstrass needed for their nineteenth-century analysis with its even greater power, together with its improved rigor and generality. [**9**, ch. 7] [**49**, p. 948] This story became so well known that it was cited by the literary critic Matthew Arnold, who wrote, "The man of genius [Newton] was continued by... completely powerless and obscure followers.... The man of intelligence [Leibniz] was continued by successors like Bernoulli, Euler, Lagrange, and Laplace—the greatest names in modern mathematics." [**1**, p. 54; cited by [**61**, p. 15]]

Now since I myself have contributed to the standard story, especially in delineating the links among Euler, Lagrange, and Cauchy, [**38**, chs. 3–6] I have a good deal of sympathy for it, but I now think that it must be modified. Maclaurin's *Treatise of Fluxions* is an important link between the calculus of Newton and Continental analysis, and Maclaurin contributed to key developments in the mathematics of his contemporaries. Let us examine the evidence for this statement.

3. THE NATURE OF MACLAURIN'S TREATISE OF FLUXIONS. Why—the standard story notwithstanding—might Maclaurin's *Treatise of Fluxions* have been able to transmit Newtonian calculus, improved and expanded, to the Continent? First, because the *Treatise of Fluxions* is not just one "Book," but two. While Book I is largely, though not entirely, geometric, Book II has a different agenda. Its title is "On the *Computations* in the Method of Fluxions." [my italics] Maclaurin began Book II by championing the power of symbolic notation in mathematics. [**63**, pp. 575–576] He explained, as Leibniz before him and Lagrange after him would agree, that the usefulness of symbolic notation arises from its generality. So, Maclaurin continued, it is important to demonstrate the rules of fluxions once

again, this time from a more algebraic point of view. Maclaurin's appreciation of the algorithmic power of algebraic and calculus notation expresses a common eighteenth-century theme, one developed further by Euler and Lagrange in their pursuit of pure analysis detached from any kind of geometric intuition. To be sure, Maclaurin, unlike Euler and Lagrange, did not wish to detach the calculus from geometry. Nonetheless, Maclaurin's second Book in fact, as well as in rhetoric, has an algorithmic character, and most of its results may be read independently of their geometric underpinnings, even if Maclaurin did not so intend. (In his Preface to Book I, he even urged readers to look at Book II before the harder parts of Book I.) [**63**, p iii] The *Treatise of Fluxions*, then, was not foreign to the Continental point of view, and may have been written in part with a Continental audience in mind.

Nor was this algebraic character a secret open only to the reader of English. There was a French translation in 1749 by the Jesuit R. P. Pézénas, including an extensive table of contents. [**62**] Lagrange, among others, seems to have used this French edition (since he cited it by the French title [**58**, p. 17] though he cited other English works in English [**58**, p. 18]). Pézénas' translation, moreover, was neither isolated nor idiosyncratic, but part of the activity of a network of Jesuits inter-
ested in mathematics and mathematical physics, especially work in English, with Maclaurin one of the authors of interest to them. [**84**, pp. 33, 221, 278, 517, 655] For instance, Pézénas himself translated other English works, including those by Desaguliers, Gardiner's logarithmic tables, and Seth Ward's *Young Mathematician's Guide* [**83**, pp. 571–2] Thus there was a well-worn path connecting English-language work with interested Continental readers. Furthermore, the two-fold character of the *Treaties of Fluxions* was noted, with special praise for Book II's treatment of series, by Silvestre-François Lacroix in the historical introduction to the second edition of his highly influential three-volume calculus textbook. [**52**, p. xxvii] Unfortunately, though, recognition of the two-fold character has been absent from the literature almost completely from Lacroix's time until the recent work by Sageng and Guicciardini. [**42**] [**78**] We shall address the reasons for this neglect in due course.

4. THE SOCIAL CONTEXT: THE SCOTTISH ENLIGHTENMENT. Another reason for doubting the standard picture comes from the social context of Maclaurin's career. Eighteenth-century Scotland, Maclaurin's home, was anything but an intellectual backwater. It was full of first-rate thinkers who energetically pursued science and philosophy and whose work was known and respected throughout Europe. One would expect Scotland's leading mathematician to share these connections and this international renown, and he did.

Although Scotland had been deprived of its independent national government by the Act of Union of 1707, it still retained, besides its independent legal system and its prevailing religion, its own educational system. The strength and energy of Scottish higher education in Maclaurin's time is owed in large part to the Scottish ruling classes, landowners and merchants alike, who saw science, mathematics, and philosophy as keys to what they called the "improvement" of their yet underdeveloped nation. [**65**, p. 254] [**80**, pp. 7–8, 10–11] [**17**, pp. 127, 132–3] Eighteenth-century Scotland, with one-tenth the population of England, had four major universities to England's two. [**80**, p. 116] Maclaurin, when he wrote the *Treatise of Fluxions*, was Professor of Mathematics at the University of Edinburgh. Edinburgh was about to become the heart of the Scottish Enlightenment, and Maclaurin until his death in 1746 was a leading figure in that city's cultural life.

Mathematics played a major role in the Scottish university curriculum. This was in part for engineers; Scottish military engineers were highly in demand even on the Continent. [**17**, p. 125] Maclaurin himself was actively interested in the applications of mathematics, and just before his untimely death had planned to write a book on the subject. [**36**] [**68**, p. xix] In addition, mathematics and Newtonian physics were part of the course of study for prospective clergyman. [**80**, p. 20] The influential "Moderate" party in the Church of Scotland appreciated the Newtonian reconciliation of science and religion. [**16**, pp. 53, 57]

Maclaurin's position in Edinburgh's cultural life was not just that of a technically competent mathematician. For instance, he was part of the Rankenian society, which met at Ranken's Tavern in Edinburgh to discuss such things as the philosophy of Bishop Berkeley; the society introduced Berkeley's philosophy to the Scottish university curriculum. [**24**, p. 222] [**17**, p. 133] [**65**, p. 197] Maclaurin and his physician friend Alexander Monro were the founders and moving spirits of the Edinburgh Philosophical Society. [**65**, p. 198] With Newton's encouragement, Maclaurin had become the chief spokesman in Scotland for the new Newtonian physics. His posthumously published book, *An Account of Sir Isaac Newton's Philosophical Discoveries*, was based on material Maclaurin used in his classes at Edinburgh, and the book was of great interest to philosophers. [**24**, p. 137] That book became well known on the Continent. It was translated into French almost as soon as it appeared, by Louis-Anne Lavirotte in 1749, and the first part appeared in Italian in Venice in 1762.

Anoher branch of Scottish science, namely medicine, also had many links with the Continent and was highly regarded there. Medical students went back and forth between Scotland, Holland, and France. [**17**, p. 135] [**80**, p. 7]

The best-known figures of eighteenth-century Scotland had major interactions with, and influence upon, Continental science and philosophy. [**39**] [**81**] Let it suffice to mention the names of four: the philosopher David Hume, who was a student at Edinburgh in Maclaurin's time; the geologist James Hutton, who attended and admired Maclaurin's lectures; [**34**, pp. 577–8] and, a bit after Maclaurin's time but still subject to his influence on Scottish higher education, the chemist Joseph Black and the economic and political philosopher Adam Smith. Maclaurin himself had twice won prizes from the Académie des Sciences in Paris, once in 1724 for a memoir on percussion, and then in 1740 (dividing the prize with Daniel Bernoulli, P. Antoine Cavalleri, and Leonhard Euler) for a memoir on the tides. [**79**, p. 611] [**39**, pp. 400–401]

Scotland in the eighteenth century nurtured first-rate intellectual work on mathematics, philosophy, science, medicine, and engineering, and did it all as part of a general European culture. [**39**, p. 412] [**81**, passim] The *Treatise of Fluxions* was the major mathematical work of a Scottish mathematician of considerable reputation on the Continent, a major work philosophically attuned to the enormously influential Newtonian physics and the Continentally popular algebraic symbolism. Such a work would certainly be of interest to Continental thinkers. Social considerations may not suffice to determine mathematical ideas, but they certainly affect the mathematician's ability to make a living, to get research support, and to promote contact and communication with other mathematicians and scientists at home and abroad. And so it was with Maclaurin.

5. MACLAURIN'S CONTINENTAL REPUTATION. An even better reason for not accepting the traditional view of Maclaurin is that his work demonstrably *was*

read in the eighteenth century, and was read by the big names of Continental mathematics. He had a Continental acquaintance through travel and correspondence. Even before the *Treatise of Fluxions*, his reputation had been enhanced by his Académie prizes and by his books on geometry. He was thus a respected member of an international network of mathematicians with interests in a wide range of subjects, and the publication of the *Treatise of Fluxions* was eagerly anticipated on the Continent.

The *Treatise of Fluxions* of 1742 was Maclaurin's major work on analysis, incorporating and somewhat dwarfing what he had done earlier. It contains an exposition of the calculus, with old results explained and many new results introduced and proved. Maclaurin seems to have included almost everything he had done in analysis and its applications to Newtonian physics. In particular, the findings of his Paris prize paper on the tides were included and expanded. His other papers, the posthumous and relatively elementary *Algebra*, and his works on geometry as such—though highly regarded—do not concern us here, but his Continental reputation was enhanced by these as well.

Let us turn now to some specific evidence for the Continental reputation of Maclaurin's major work. In 1741, Euler wrote to Clairaut that, though he had not yet seen the Paris prize papers on the tides, "from Mr. Maclaurin I expect only excellent ideas." [**47**, p. 87] Euler added that he had heard from England (presumably from his correspondent James Stirling) that Maclaurin was bringing out a book on "differential calculus," and asked Clairaut to keep him posted about this. In turn, Clairaut asked Maclaurin later in 1741 about his plans for the book, [**66**, p. 348] which Clairaut wanted to see before publishing his own work on the shape of the earth. [**47**, p. 110] Euler did get the *Treatise of Fluxions*, and read enough of it quickly to praise it in a letter to Goldbach in 1743. [**48**, p. 179] Jean d'Alembert, in his *Traité de dynamique* of 1743, [**22**, sec. 37, n.] praised the rigor brought to calculus by the *Treatise of Fluxions*. D'Alembert's most recent biographer, Thomas Hankins, argues that Maclaurin's *Treatise*, appearing at this time, helped persuade d'Alembert that gravity could best be described as a continuous acceleration rather than a series of infinitesimal leaps. [**44**, p. 167] D'Alembert's general approach to the foundations of the calculus in terms of limits clearly was influenced by Newton's and Maclaurin's championing of limits over infinitesimals, in particular by Maclaurin's clear description of limits in one of the parts of his *Treatise of Fluxions* that explicitly responds to Berkeley's objections (and which incidentally may be the first explicit description of the tangent as the limit of secant lines; see Section 7). [**44**, p. 23] [**63**, pp. 422–3] Lagrange in his *Analytical Mechanics* [**55**, p. 243] said that Maclaurin, in the *Treatise of Fluxions*, was the first to treat Newton's laws of motion in the language of the calculus in a coordinate system fixed in space. Though C. Truesdell [**80**, pp. 250–3] has shown that Lagrange was wrong because Johann Bernoulli and Euler were ahead of Maclaurin on this, the fact that Lagrange believed this is one more piece of evidence for the Continental reputation of Maclaurin as mathematician and physicist.

6. MACLAURIN'S MATHEMATICS AND ITS IMPORTANCE. The previous points show that Maclaurin could have been influential, but not that he was. Five examples will reveal both the nature of Maclaurin's techniques and the scope of his influence: a special case of the Fundamental Theorem of Calculus; Maclaurin's treatment of maxima and minima for functions of one variable; the attraction of spheroids; what is now called the Euler-Maclaurin summation formula; and elliptic integrals.

a. Key Methods in the Calculus. Two methods were central to the study of real-variable calculus in the eighteenth and nineteenth centuries. One of these is studying real-valued functions by means of power-series representations. This tradition is normally thought first to flower with Euler; it is then most closely associated with Lagrange, and, later for complex variables, with Weierstrass. The second such method is that of basing the foundations of the calculus on the algebra of inequalities—what we now call delta-epsilon proof techniques—and using algebraic inequalities to prove the major results of the calculus; this tradition is most closely associated with the work of Cauchy in the 1820's. I have traced these traditions back to Lagrange and Euler in my work on the origins of Cauchy's calculus. [**38**, chs. 3–6] It is surprising, at least if one accepts the standard picture of the history of the calculus, that both of these methods—studying functions by power series, basing foundations on inequalities—were materially advanced by Maclaurin in the *Treatise of Fluxions*. It is especially striking that the importance of Maclaurin's work on series—work based, it is well to remember, on Newton's use of infinite series—was recognized and praised in 1810 by Lacroix, who also linked it with the series-based calculus of Lagrange. [**52**, p. xxxiii]

Maclaurin skillfully used algebraic inequalities in his proof of a special case of the Fundamental Theorem of Calculus. He showed, for a particular function, that if one takes the fluxion of the area under the curve whose equation is $y = f(x)$, one gets the function $f(x)$. In his proof, Maclaurin adapted the intuition underlying Newton's argument for this fact in *De Analysi* [**69**]—that the rate of change of the area under a curve is measured by the height of the curve—but Maclaurin's proof is more rigorous. Although Maclaurin's argument proceeds algebraically, the concepts involved resemble those of the Greek "method of exhaustion" (more precisely termed by Dijksterhuis "indirect passage to the limit"). [**26**, p. 130] A key step in this Greek work is first to assume that two equal areas or expressions for areas are *unequal*, and then to argue to a contradiction by using inequalities that hold among various rectilinear areas. Newton in the *Principia* had based proofs of new results about areas and curves on methods akin to those of the Greeks. Maclaurin carried this much further. It was Maclaurin's "conservative" allegiance to Archimedean *geometric* methods that led him to buttress the *kinematic* intuition of Newton's calculus with *algebraic* inequality proofs.

What Maclaurin proved in the example under discussion is that, if the area under a curve up to x is given by x^n, the ordinate of the curve must be $y = nx^{n-1}$, which is known to be the fluxion of x^n. [**63**, pp. 752–754] Maclaurin's diagram for this is much like the one Newton gave in the *DeAnalysi*. [**69**, pp. 3–4] Maclaurin

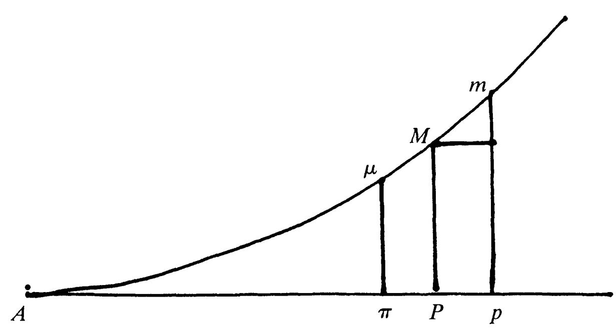


Figure 1

began by saying that, since x and y increase together, the following inequality holds between the areas shown:

$$x^n - (x-h)^n < yh < (x+h)^n - x^n. \tag{1}$$

(Maclaurin gave this inequality verbally; I have supplied the "$<$" signs; also, I use "h" for the increment where Maclaurin used "o".) Now Maclaurin recalled an algebraic identity he had proved earlier: [**63**, p. 583; inequality notation added]

$$\text{If } E < F, \text{ then } nF^{n-1}(E-F) < E^n - F^n < nE^{n-1}(E-F). \tag{2}$$

(It may strike the modern reader that, since nx^{n-1} is the derivative of x^n, this second inequality is a special case of the mean-value theorem for derivatives. I shall return to this point later.)

Now, letting $x-h$ play the role of F and x play the role of E, $E-F$ is h and the first inequality in (2) yields

$$n(x-h)^{n-1}h < x^n - (x-h)^n.$$

Similarly, if $F = x$ and $E = x+h$, then $E - F = h$ and the second inequality in (2) becomes

$$(x+h)^n - x^n < n(x+h)^n h.$$

Combining these with inequality (1) about the areas, Maclaurin obtained

$$n(x-h)^{n-1}h < yh < n(x+h)^{n-1}h.$$

Dividing by h produces

$$n(x-h)^{n-1} < y < n(x+h)^{n-1}. \tag{3}$$

Recall that, given that the area was x^n, Maclaurin was seeking an expression for y, the fluxion of that area. A modern reader, having reached the inequality (3), might stop, perhaps saying "let h go to zero, so that y becomes nx^{n-1}," or perhaps justifying the conclusion by appealing to the delta-epsilon characterization of limit. What Maclaurin did instead was what Archimedes might have done, a double reductio ad absurdum. But what Archimedes might have done geometrically and verbally, Maclaurin did algebraically. He assumed first that y is *not* equal to nx^{n-1}. Then, he said, it must be equal to $nx^{n-1} + r$ for some r. First, he considered the case when this r was positive. This will lead to a contradiction if h is chosen so that $y = n(x+h)^{n-1}$, since, he observed, inequality (3) will be violated when $h = (x^{n-1} + r/n)^{1/(n-1)}$. Similarly, he calculated the h that produces a contradiction when r is assumed to be negative. Thus there can be no such r, and $y = nx^{n-1}$. [**63**, p. 753]

Maclaurin introduced this proof by saying something surprising for a *Treatise of Fluxions*: that the use of the inequalities makes the demonstration of the value of y "independent of the notion of a fluxion." [**63**, p. 752] (Of course one would need the notion of fluxion to interpret y as the fluxion of the area function x^n, but the proof itself is algebraic.) This proof was presumably part of his agenda in writing the more algebraic Book II of the *Treatise* for an audience on the Continent, where fluxions were suspect as involving the idea of motion. Later Lagrange, in seeking his purely algebraic foundation for the calculus, explicitly said he wanted to free the calculus from fluxions and what he called the "foreign idea" of motion. It is thus striking that Lagrange's *Théorie des fonctions analytiques* (1797) gives a more general version of the kind of argument Maclaurin had given, applying to *any* increasing function that satisfies the geometric inequality expressed in (1). In

place of the algebraic inequality (2), Lagrange used the mean-value theorem. [**58**, pp. 238–9] [**38**, pp. 156–158] The similarity of the two arguments does not prove influence, of course, but it certainly demonstrates that Maclaurin's work, which we know Lagrange read (e.g., [**58**, p. 17]), uses the algebra of inequalities in a way consistent with that used by Lagrange and his successors.

Maclaurin's argument exemplifies the way his *Treatise* reconciles the old and the new. The double *reductio ad absurdum* reflects his Archimedean agenda. Treating the area as generated by a moving vertical line, and then searching for the relationship between the area and its fluxion, are Newtonian. Maclaurin did not have a general proof of the Fundamental Theorem in this argument, but relied on an inequality based on the specific properties of a specific function. Nonetheless, he had the precise bounding inequalities for the area function used later by Lagrange, and he used an algebraic inequality proof in a manner that would not disgrace a nineteenth-century analyst.

Inequality-based arguments in the calculus as used by Lagrange and Cauchy owe a lot to the eighteenth-century study of algebraic approximations, and it once seemed to me that this was their origin. But the algebra of inequalities as used in Continental analysis, especially in d'Alembert's pioneering treatment of the tangent as the limit of secants in the article "Différentiel" in the *Encyclopédie*, [**19**] must owe something also to Maclaurin's translation of Archimedean geometry into algebraic dress to justify results in calculus. Throughout the eighteenth century, practitioners of the limit tradition on the Continent use inequalities; a clear line of influence connects Maclaurin's admirer d'Alembert, Simon L'Huilier (who was a foreign member of the Royal Society), the textbook treatment of limits by Lacroix, and, finally, Cauchy. [**38**, pp. 80–87]

Now let us turn to some of Maclaurin's work on series. There is, of course, the Maclaurin series, that is, the Taylor series expanded around zero. This result Maclaurin himself credited to Taylor, and it was known earlier to Newton and Gregory. It was called the Maclaurin series by John F. W. Herschel, Charles Babbage, and George Peacock in 1816 [**51**, pp. 620–21] and by Cauchy in 1823. [**14**, p. 257] Since it was obvious that Maclaurin had not invented it, the attribution shows appreciation by these later mathematicians for the way Maclaurin used the series to study functions. A key application is Maclaurin's characterization of maxima, minima, and points of inflection of an infinitely differentiable function by means of its successive derivatives. When the first derivative at a point is zero, there is a maximum if the second derivative is negative there, a minimum if it is positive. If the second derivative is also zero, one looks at higher derivatives to tell whether the point is a maximum, minimum, or point of inflection. These results can be proved by looking at the Taylor series of the function near the point in question, and arguing on the basis of the inequalities expressed in the definition of maximum and minimum. For instance (in modern [Lagrangian] notation), if $f(x)$ is a maximum, then

$$f(x) > f(x+h) = f(x) + hf'(x) + h^2/2!f''(x) + \cdots, \text{ and} \qquad (4)$$
$$f(x) > f(x-h) = f(x) - hf'(x) + h^2/2!f''(x) - \cdots$$

if h is small. If the derivatives are bounded, and if h is taken sufficiently small so that the term in h dominates the rest, the inequalities (4) can both hold only if $f'(x) = 0$. If $f'(x) = 0$, then the h^2 term dominates, and the inequalities (4) hold only if $f''(x)$ is negative. And so on.

I have traced Cauchy's use of this technique back to Lagrange, and from Lagrange back to Euler. [**38**, pp. 117–118] [**37**, pp. 157–159] [**58**, pp. 235–6] [**29**,

Secs. 253–254] But this technique is explicitly worked out in Maclaurin's *Treatise of Fluxions*. Indeed, it appears twice: once in geometric dress in Book I, Chapter IX, and then more algebraically in Book II. [**63**, pp. 694–696] Euler, in the version he gave in his 1755 textbook, [**20**] does not refer to Maclaurin on this point, but then he makes few references in that book at all. Still we might suspect, especially knowing that Stirling told Euler in a letter of 16 April 1738 [**91**] that Maclaurin had some interesting results on series, that Euler would have been particularly interested in looking at Maclaurin's applications of the Taylor series. Certainly Lacroix's praise for Maclaurin's work on series must have taken this set of results into account. [**52**, p. xxvii] Even more important, Lagrange, in unpublished lectures on the calculus from Turin in the 1750's, after giving a very elementary treatment of maxima and minima, referred to volume II of Maclaurin's *Treatise of Fluxions* as the chief source for more information on the subject. [**7**, p. 154] Since Lagrange did not mention Euler in this connection at all, Lagrange could well have not even have seen the *Institutiones calculi differentialis* of 1755 when he made this reference. This Taylor-series approach to maxima and minima (with the Lagrange remainder supplied for the Taylor series) plays a major role in the work of Lagrange, and later in the work of Cauchy. It is because Maclaurin thought of maxima and minima, and of convexity and concavity, in Archimedean geometrical terms that he was led to look at the relevant inequalities, just as the geometry of Archimedes helped Maclaurin formulate some of the inequalities he used to prove his special case of the Fundamental Theorem of Calculus.

b. Ellipsoids. We now turn to work in applied mathematics that constitutes one of Maclaurin's great claims to fame: the gravitational attraction of ellipsoids and the related problem of the shape of the earth. Maclaurin is still often regarded as the creator of the subject of attraction of ellipsoids. [**85**, pp. 175, 374] In the eighteenth century, the topic attracted serious work from d'Alembert, A.-C. Clairaut, Euler, Laplace, Lagrange, Legendre, Poisson, and Gauss. In the twentieth century, Subramanyan Chandrasekhar (later Nobel laureate in physics) devoted an entire chapter of his classic *Ellipsoidal Figures of Equilibrium* to the study of Maclaurin spheroids (figures that arise when homogeneous bodies rotate with uniform angular velocity), the conditions of stability of these spheroids and their harmonic modes of oscillation, and their status as limiting cases of more general figures of equilibrium. Such spheroids are part of the modern study of classical dynamics in the work of scientists like Chandrasekhar, Laurence Rossner, Carl Rosenkilde, and Norman Lebovitz. [**15**, pp. 77–100] Already in 1740 Maclaurin had given a "rigorously exact, geometrical theory" of homogeneous ellipsoids subject to inverse-square gravitational forces, and had shown that an oblate spheroid is a possible figure of equilibrium under Newtonian mutual gravitation, a result with obvious relevance for the shape of the earth. [**39**, p. 172] [**86**, p. xix] [**85**, p. 374]

Of particular importance was Maclaurin's decisive influence on Clairaut. Maclaurin and Clairaut corresponded extensively, and Clairaut's seminal 1743 book *La Figure de la Terre* [**18**] frequently, explicitly, and substantively cites his debts to Maclaurin's work. [**39**, pp. 590–597] A key result, that the attractions of two confocal ellipsoids at a point external to both are proportional to their masses and are in the same direction, was attributed to Maclaurin by d'Alembert, an attribution repeated by Laplace, Lagrange, and Legendre, then by Gauss, who went back to Maclaurin's original paper, and finally by Lord Kelvin, who called it "Maclaurin's splendid theorem." [**15**, p. 38] [**85**, pp. 145, 409] Lagrange began his own memoir on the attraction of ellipsoids by praising Maclaurin's treatment in the

prize paper of 1740 as a masterwork of geometry, comparing the beauty and ingenuity of Maclaurin's work with that of Archimedes, [**57**, p. 619] though Lagrange, typically, then treated the problem analytically. Maclaurin's eighteenth- and nineteenth-century successors also credit him with some of the key methods used in studying the equilibrium of fluids, such as the method of balancing columns. [**39**, p. 597] Maclaurin's work on the attraction of ellipsoids shows how his geometric insights fruitfully influenced a subject that later became an analytic one.

c. The Euler-Maclaurin Formula. The Euler-Maclaurin formula expresses the value of definite integrals by means of infinite series whose coefficients involve what are now called the Bernoulli numbers. The formula shows how to use integrals to find the partial sums of series. Maclaurin's version, in modern notation, is:

$$\sum_{h=0}^{\infty} F(a+h) = \int_0^a F(x)\,dx + 1/2F(a) + 1/2F'(a) - 1/720F'''(a) + 1/30240F^{(v)}(a) - \cdots$$

[**35**, pp. 84–86]

James Stirling in 1738, congratulating Euler on his publication of that formula, told Euler that Maclaurin had already made it public in the first part of the *Treatise of Fluxions*, which was printed and circulating in Great Britain in 1737. [**47**, p. 88n] [**91**, p. 178] (On this early publication, see also [**63**, pp. iii, 691n]). P. L. Griffiths has argued that this simultaneous discovery rests on De Moivre's work on summing reciprocals, which also involves the so-called Bernoulli numbers. [**40**] [**41**, pp. 16–17] [**25**, p. 19] In any case, Euler and Maclaurin derived the Euler-Maclaurin formula in essentially the same way, from a similar geometric diagram and then by integrating various Taylor series and performing appropriate substitutions to find the coefficients. [**31**] [**32**] [**33**] Maclaurin's approach is no more Archimedean or geometric than Euler's; they are similar and independent. [**63**, pp. 289–293, 672–675] [**35**, pp. 84–93] [**67**] In subsequent work, Euler went on to extend and apply the formula further to many other series, especially in his *Introductio in analysin infinitorum* of 1748 and *Institutiones calculi differentialis* of 1755. [**35**, p. 127] But Maclaurin, like Euler, had applied the formula to solve many problems. [**63**, pp. 676–693] For instance, Maclaurin used it to sum powers of arithmetic progressions and to derive Stirling's formula for factorials. He also derived what is now called the Newton-Cotes numerical integration formula, and obtained what is now called Simpson's rule as a special case. It is possible that his work helped stimulate Euler's later, fuller investigations of these important ideas.

In 1772, Lagrange generalized the Euler-Maclaurin formula, which he obtained as a consequence of his new calculus of operators. [**53**] [**35**, pp. 169, 261] In 1834, Jacobi provided the formula with its remainder term, [**46**, pp. 263, 265] in the same paper in which he first introduced what are now called the Bernoulli polynomials. Jacobi, who called the result simply the Maclaurin summation formula, cited it directly from the *Treatise of Fluxions*. [**46**, p. 263] Later, Karl Pearson used the formula as an important tool in his statistical work, especially in analyzing frequency curves. [**72**, pp. 217, 262]

The Euler-Maclaurin formula, then, is an important result in the mainstream of mathematics, with many applications, for which Maclaurin, both in the eighteenth century and later on, has rightly shared the credit.

d. Elliptic Integrals. Some integrals (Maclaurin used the Newtonian term "fluents"), are algebraic functions, Maclaurin observed. Others are not, but some of these can be reduced to finding circular arcs, others to finding logarithms. By analogy, Maclaurin suggested, perhaps a large class of integrals could be studied by being reduced to finding the length of an elliptical or hyperbolic arc. [**63**, p. 652] By means of clever geometric transformations, Maclaurin was able to reduce the integral that represented the length of a hyperbolic arc to a 'nice' form. Then, by algebraic manipulation, he could reduce some previously intractable integrals to that same form. His work was translated into analysis by d'Alembert and then generalized by Euler. [**13**, p. 846] [**23**] [**27**, p. 526] [**28**, p. 258] In 1764, Euler found a much more elegant, general, and analytic version of this approach, and worked out many more examples, but cited the work of Maclaurin and d'Alembert as the source of his investigation. A.-M. Legendre, the key figure in the eighteenth-century history of elliptic integrals, credited Euler with seeing that, by the aid of a good notation, arcs of ellipses and other transcendental curves could be as generally used in integration as circular and logarithmic arcs. [**45**, p. 139] Legendre was, of course, right that "elliptic integrals" encompass a wide range of examples; this was exactly Maclaurin's point. Thus, although his successors accomplished more, Maclaurin helped initiate a very important investigation and was the first to appreciate its generality. Maclaurin's geometric insight, applied to a problem in analysis, again brought him to a discovery.

7. OTHER EXAMPLES OF MACLAURIN'S MATHEMATICAL INFLUENCE. The foregoing examples provide evidence of direct influence of the *Treatise of Fluxions* on Continental mathematics. There is much more. For instance, Lacroix, in his treatment of integrals by the method of partial fractions, called it "the method of Maclaurin, followed by Euler." [**52**, Vol. II, p. 10] [**63**, pp. 634–644] Of interest too is Maclaurin's clear understanding of the use of limits in founding the calculus, especially in the light of his likely influence on d'Alembert's treatment of the foundations of the calculus by means of limits in the *Encyclopédie*, which in turn influenced the subsequent use of limits by L'Huilier, Lacroix, and Cauchy, [**38**, chapter 3] (and on Lagrange's acceptance of the limit approach in his early work in the 1750's). [**7**] Although the largest part of Maclaurin's reply to Berkeley was the extensive proof of results in calculus using Greek methods, he was willing to explain important concepts using limits also. In particular, Maclaurin wrote, "As the tangent of an arch [arc] is the right line that limits the position of all the secants that can pass through the point of contact... though strictly speaking it be no secant; so a ratio may limit the variable ratios of the increments, though it cannot be said to be the ratio of any real increments." [**63**, p. 423] Maclaurin's statement answers Berkeley's chief objection—that the increment in a function's value is first treated as non-zero, then as zero, when one calculates the limit of the ratio of increments or finds the tangent to a curve. Maclaurin's statement is in the tradition of Newton's *Principia* (Book I, Scholium to Lemma XI), but is in a form much closer to the later work of d'Alembert on secants and tangents. [**20**] Maclaurin pointed out that most of the propositions of the calculus that he could prove by means of geometry "may be *briefly* demonstrated by this method [of limits]." [**63**, p. 87, my italics]

In addition, Maclaurin had considerable influence in Britain, on mathematicians like John Landen (whose work on series was praised by Lagrange), Robert Woodhouse (who sparked the new British interest in Continental work about 1800), and on Edward Waring and Thomas Simpson, whose names are attached to

results well known today. [**42**] Going beyond the calculus, Maclaurin's purely geometric treatises were read and used by French geometers of the stature of Chasles and Poncelet. [**90**, p. 145] Thus, though Maclaurin may not have been the towering figure Euler was, he was clearly a significant and respected mathematician, and the *Treatise of Fluxions* was far more than an unread tome whose weight served solely to crush Bishop Berkeley.

8. WHY A TREATISE OF FLUXIONS? The *Treatise of Fluxions* was not really intended as a reply to Berkeley. Maclaurin could have refuted Berkeley with a pamphlet. It was not a student handbook either; this work is far from elementary. Nor was it merely written to glory in Greek geometry. Maclaurin wrote several works on geometry per se. But he was no antiquarian. Instead, the *Treatise of Fluxions* was the major outlet for Maclaurin's solution of significant research problems in the field we now call analysis. Geometry, as the examples I gave illustrate, was for Maclaurin a source of motivation, of insight, and of problem-solving power, as well as being his model of rigor.

For Maclaurin, rigor was not an end in itself, or a goal pursued for purely philosophical reasons. It was motivated by his research goals in analysis. For instance, Maclaurin developed his theory of maxima, minima, points of inflection, convexity and concavity, orders of contact, etc., because he wanted to study curves of all types, including those that cross over themselves, loop around and are tangent to themselves, and so on. He needed a sophisticated theory to characterize the special points of such curves. Again, in problems as different as studying the attraction of ellipsoids and evaluating integrals approximately, he needed to use infinite series and know how close he was to their sum. Thus, rigor, to Maclaurin, was not merely a tool to defend Newton's calculus against Berkeley—though it was that—nor just a response to the needs of a professor to present his students a finished subject—though it may have been that as well. In many examples, Maclaurin's rigor serves the needs of his research.

Moreover, the *Treatise of Fluxions* contains a wealth of applications of fluxions, from standard physical problems such as curves of quickest descent to mathematical problems like the summation of power series—in the context of which, incidentally, Maclaurin gave what may be the earliest clear definition of the sum of an infinite series: "There are progressions of fractions which may be continued at pleasure, and yet the sum of the terms be always less than a certain finite number. If the difference betwixt their sum and this number decrease in such a manner, that by continuing the progression it may become less than any fraction how small soever that can be assigned, this number is the *limit of the sum of the progression*, and is what is understood by the value of the progression when it is supposed to be continued indefinitely." [**63**, p. 289] Thus, though eighteenth-century Continental mathematicians did not care passionately about foundations, [**38**, pp. 18–24] they could still appreciate the *Treatise of Fluxions* because they could mine it for results and techniques.

9. WHY THE TRADITIONAL VIEW? If the reader is convinced by now that the traditional view is wrong, that Maclaurin's Treatise did not mark the end of the Newtonian tradition, and that not all of modern analysis stems solely from the work of Leibniz and his school, the question arises, how did that traditional view come to be, and why it has been so persistent?

Perhaps the traditional view could be explained as follows. Consider the approach to mathematics associated with Descartes: symbolic power, not debates

over foundations; problem-solving power, not axioms or long proofs. The Cartesian approach to mathematics is clearly reflected in the work and in the rhetoric of Leibniz, Johann Bernoulli, Euler, Lagrange—especially in the historical prefaces to his influential works—and even Cauchy. These men, the giants of their time, are linked in a continuous chain of teachers, close colleagues, and students. Some topics, like partial differential equations and the calculus of variations, were developed mostly on the Continent. Moreover, the Newton-Leibniz controversy helped drive English and Continental mathematicians apart. Thus the Continental tradition can be viewed as self-contained, and the outsider sees no need for eighteenth-century Continental mathematicians to struggle through 750 pages of a *Treatise of Fluxions*, which is at best in the Newtonian notation and at worst in the language of Greek geometry. Lagrange's well-known boast that his *Analytical Mechanics* [**55**] had (and needed) no diagrams, thus opposing analysis to geometry at the latter's expense, reinforced these tendencies and enshrined them in historical discourse. But the explanation we have just given does not suffice to explain the strength, and persistence into the twentieth century, of the standard interpretation. The traditional view of Maclaurin's lack of importance has been reinforced by some other historiographical tendencies that deserve our critical attention.

The traditional picture of Maclaurin's *Treatise of Fluxions* radically separates his work on foundations, which it regards as geometric, sterile, and antiquarian, from his important individual results, which often are mentioned in histories of mathematics but are treated in isolation from the purpose of the *Treatise*, in isolation from one another, and in isolation from Maclaurin's overall approach to mathematics. Strangely, both externalist and internalist historians, each for different reasons, have reinforced this picture.

For instance, in the English-speaking world, viewing the *Treatise* as only about Maclaurin's foundation for the calculus, and thus as a dead end, has been perpetuated by the "decline of science in England" school of the history of eighteenth-century science, stemming from such early nineteenth-century figures as John Playfair, and, especially, Charles Babbage. [**77**] [**2**] [**4**] Babbage felt strongly about this because he was a founder of the Cambridge Analytical Society, which fought to introduce Continental analysis into Cambridge in the early nineteenth century. This group had an incentive to exaggerate the superiority of Continental mathematics and downgrade the British, as is exemplified by their oft-quoted remark that the principles of "pure d-ism" should replace what they called the "dot-age" of the University. [**5**, ch. 7] [**10**, p. 274] The pun, playing on the Leibnizian and Newtonian notation in calculus, may be found in [**2**, p. 26]. These views continued to be used in the attempt by Babbage and others to reform the Royal Society and to increase public support for British science.

It is both amusing and symptomatic of the misunderstanding of Maclaurin's influence that Lacroix's one-volume treatise on the calculus of 1802, [**50**] translated into English by the Cambridge Analytical Society with added notes on the method of series of Lagrange, [**51**] was treated by them, and has been considered since, as a purely "Continental" work. But Lacroix's short treatise was based on the concept of limit, which was Newtonian, elaborated by Maclaurin, adapted by d'Alembert and L'Huilier, and finally systematized by Lacroix. [**38**, pp. 81–86] Moreover, the translators' notes by Babbage, Herschel, and Peacock supplement the text by studying functions by their Taylor series, thus using the approach that Lacroix himself, in his multi-volume treatise of 1810, had attributed to Maclaurin. This is, of course, not to deny the overwhelming importance of the contributions of Euler and Lagrange, both to the mathematics taught by the Analytical Society and to

that included by Lacroix in his 1802 book, nor to deny the Analytical Society's emphasis on a more abstract and formal concept of function. But all the same, Babbage, Herschel, and Peacock were teaching some of Maclaurin's ideas without realizing this.

In any case, the views expressed by Babbage and others have strongly influenced Cambridge-oriented writers like W. W. Rouse Ball, who said that the history of eighteenth-century English mathematics "leads nowhere." [**5**, p. 98] H. W. Turnbull, though he wrote sympathetically about Maclaurin's mathematics on one occasion, [**88**] blamed Maclaurin on another occasion for the decline: "When Maclaurin produced a great geometrical work on fluxions, the scale was so heavily loaded that it diverted England from Continental habits of thought. During the remainder of the century, British mathematics were relatively undistinguished." [**89**, p. 115]

Historians of Scottish thought, working from their central concerns, have also unintentionally contributed to the standard picture. George Elder Davie, arguing from social context to a judgment of Maclaurin's mathematics, held that the Scots, unlike the English, had an anti-specialist intellectual tradition, based in philosophy, and emphasizing "cultural and liberal values." Wishing to place Maclaurin in this context, Davie stressed what he called Maclaurin's "mathematical Hellenism," [**24**, p. 112] and was thus led to circumscribe the achievement of the *Treatise of Fluxions* as having based the calculus "on the Euclidean foundations provided by [Robert] Simson," [**24**, p. 111] who had made the study of the writings of the classical Greek geometers the "national norm" in Scotland. The "Maclaurin is a geometer" interpretation among Scottish historians has been further reinforced by a debate in 1838 over who would fill the Edinburgh chair in mathematics. Phillip Kelland, a candidate from Cambridge, was seen as the champion of Continental analysis, while the partisans of Duncan Gregory argued for a more geometrical approach. Wishing to enlist the entire Scottish geometric tradition on the side of Gregory, Sir William Hamilton wrote, "The great Scottish mathematicians, ... even Maclaurin, were decidedly averse from the application of the mechanical procedures of algebra." [**24**, p. 155] Though Kelland eventually won the chair, the dispute helped spread the view that Maclaurin had been hostile to analysis. More recently, Richard Olson has characterized Scottish mathematics after Maclaurin as having been conditioned by Scottish common-sense philosophy to be geometric in the extreme. [**70**, pp. 4, 15] [**71**, p. 29] But in emphasizing Maclaurin's influence on this development, Olson, like Davie, has overstated the degree to which Maclaurin's approach was geometric.

By contrast, consider internalist historians. The treatment of Maclaurin's results as isolated reflects what Herbert Butterfield called the Whig approach to history, viewing the development of eighteenth-century mathematics as a linear progression toward what we value today, the collection of results and techniques which make up classical analysis. Thus, mathematicians writing about the history of this period, from Moritz Cantor in the nineteenth century to Hermann Goldstine and Morris Kline in the twentieth, tell us what Maclaurin did with specific results, some named after him, for which they have mined the *Treatise of Fluxions*. [**13**, pp. 655–63] [**35**, pp. 126ff, 167–8] [**49**, pp. 522–3, 452, 442] They either neglect the apparently fruitless work on foundations, or, viewing it as geometric, see it as a step backward. It is of course true that many Continental mathematicians used Maclaurin's results without accepting the geometrical and Newtonian insights that Maclaurin used to produce them. But without those points of view, Maclaurin would not have produced those results.

Both externalist and internalist historians, then, have treated Maclaurin's work in the same way: as a throwback to the Greeks, with a few good results that happen to be in there somewhat like currants in a scone. Further, the fact that Maclaurin's book, especially its first hundred pages, is very hard to read, especially for readers schooled in modern analysis, has encouraged historians who focus on foundations to read only the introductory parts. The fact that there is so much material has encouraged those interested in results to look only at the sections of interest to them. And the fact that the first volume is so overwhelmingly geometric serves to reinforce the traditional picture once again whenever anybody opens the *Treatise*. The recent Ph.D. dissertation by Erik Sageng [**78**] is the first example of a modern scholarly study of Maclaurin's *Treatise* in any depth. The standard picture has not yet been seriously challenged in print.

10. SOME FINAL REFLECTIONS. Maclaurin's work had Continental influence, but with an important exception—his geometric foundation for the calculus. Mastering this is a major effort, and I know of no evidence that any eighteenth-century Continental mathematician actually did so. Lagrange perhaps came the closest. In the introduction to his *Théorie des fonctions analytiques*, Lagrange could say only, Maclaurin did a good job basing calculus on Greek geometry, so it can be done, but it is very hard. [**58**, p. 17] In an unpublished draft of this introduction, Lagrange said more pointedly: "I appeal to the evidence of all those with the courage to read the learned treatise of Maclaurin and with enough knowledge to understand it: have they, finally, had their doubts cleared up and their spirit satisfied?" [**73**, p. 30]

Something else may have blunted people's views of the mathematical quality of Maclaurin's *Treatise*. The way the book is constructed partly reflects the Scottish intellectual milieu. The Enlightenment in Britain, compared with that on the Continent, was marked less by violent contrast and breaks with the past than by a spirit of bridging and evolution. [**75**, pp. 7–8, 15] Similarly, Scottish reformers operated less by revolution than by the refurbishment of existing institutions. [**16**, p. 8] These trends are consistent with the two-fold character of the *Treatise of Fluxions*: a synthesis of the old and the new, of geometry and algebra, of foundations and of new results, a refurbishment of Newtonian fluxions to deal with more modern problems. This contrasts with the explicitly revolutionary philosophy of mathematics of Descartes and Leibniz, and thus with the spirit of the *mathématicien* of the eighteenth century on the Continent.

Of course Scotland was not unmarked by the conflicts of the century. During the Jacobite rebellion in 1745, Maclaurin took a major role in fortifying Edinburgh against the forces of Bonnie Prince Charlie. When the city was surrendered to the rebels, Maclaurin fled to York. Before his return, he became ill, and apparently never really recovered. He briefly resumed teaching, but died in 1746 at the relatively young age of forty-eight. Nonetheless, the Newtonian tradition in the calculus was not a dead end. Maclaurin in his lifetime, and his *Treatise of Fluxions* throughout the century, transmitted an expanded and improved Newtonian calculus to Continental analysts. And Maclaurin's geometric insight helped him advance analytic subjects.

We conclude with the words of an eighteenth-century Continental mathematician whose achievements owe much to Maclaurin's work. [**39**, pp. 172, 412–425, 590–597] The quotation [**66**, p. 350] illustrates Maclaurin's role in transmitting the Newtonian tradition to the Continent, the respect in which he was held, and the eighteenth-century social context essential to understanding the fate of his work.

In 1741, Alexis-Claude Clairaut wrote to Colin Maclaurin, "If Edinburgh is, as you say, one of the farthest corners of the world, you are bringing it closer by the number of beautiful discoveries you have made."

ACKNOWLEDGMENT. I thank the Department of History and Philosophy of Science of the University of Leeds, England, for its hospitality while I was doing much of this research, and the Mathematics Department of the University of Edinburgh, where I finished it. I also thank Professor G. N. Cantor for material as well as intellectual assistance, and Professors J. R. R. Christie and M. J. S. Hodge for stimulating and valuable conversations.

REFERENCES

1. Arnold, Matthew, The Literary Importance of Academies, in Matthew Arnold, *Essays in Criticism*, Macmillan, London, 1865, 42–79. Cited in [**61**].
2. Babbage, Charles, *Passages from the Life of a Philosopher*, Longman, London, 1864.
3. [Babbage, Charles], "Preface" to *Memoirs of the Analytical Society* Cambridge, J. Smith, 1813. Attributed to Babbage by Anthony Hyman, *Charles Babbage: Pioneer of the Computer*, Princeton University Press, Princeton, 1982.
4. Babbage, Charles, *Reflections on the Decline of Science in England, and Some of its Causes*, B. Fellowes, London, 1830.
5. Ball, W. W. Rouse, *A History of the Study of Mathematics at Cambridge*, Cambridge University Press, Cambridge, 1889.
6. Berkeley, George, *The Analyst, or a Discourse Addressed to an Infidel Mathematician*, in A. A. Luce and T. R. Jessop, eds., *The Works of George Berkeley*, vol. 4, T. Nelson, London, 1951, 65–102.
7. Borgato, Maria Teresa, and Luigi Pepe, Lagrange a Torino (1750–1759) e le sue lezioni inedite nelle R. Scuole di Artiglieria, *Bollettino di Storia delle Scienze Matematiche*, 1987, 7: 3–180.
8. Bourbaki, Nicolas, *Elements d'histoire des mathématiques*, Paris Hermann, Paris, 1960.
9. Boyer, Carl, *The History of the Calculus and Its Conceptual Development*, Dover, New York, 1959.
10. Cajori, Florian, *A History of the Conceptions of Limits and Fluxions in Great Britain from Newton to Woodhouse*, Open Court, Chicago and London, 1919.
11. Cajori, Florian, *A History of Mathematics*, 2d. ed., Macmillan, New York, 1922.
12. Cantor, G. N., Anti-Newton, in J. Fauvel et al., eds., *Let Newton Be!*, Oxford University Press, Oxford, 1988, pp. 203–222.
13. Cantor, Moritz, *Vorlesüngen über Geschichte der Mathematik*, vol. 3, Teubner, Leipzig, 1898.
14. Cauchy, A.-L., *Résumé des leçons données à l'école royale polytechnique sur le calcul infinitésimal*, in *Oeuvres complètes*, Ser. 2, vol. 4, Gauthier-Villars, Paris, 1899.
15. Chandrasekhar, S., *Ellipsoidal Figures of Equilibrium*, Yale, New Haven, 1969.
16. Chitnis, Anand, *The Scottish Enlightenment: A Social History*, Croom Helm, London, 1976.
17. Christie, John R. R., The Origins and Development of the Scottish Scientific Community, 1680–1760, *History of Science* 12 (1974), 122–141.
18. Clairaut, A.-C., *Théorie de la figure de la terre*, Duraud, Paris, 1743.
19. d'Alembert, Jean, "Différentiel," in [**21**].
20. d'Alembert, Jean, and de la Chapelle, "Limite," in [**21**].
21. d'Alembert, Jean, et al., eds., *Dictionnaire encyclopédique des mathématiques*, Hotel de Thou, Paris, 1789, which collects the mathematical articles from the Diderot-d'Alembert *Encyclopédie*.
22. d'Alembert, Jean, *Traité de dynamique*, David l'Aîné, Paris, 1743.
23. d'Alembert, Jean, Récherches sur le calcul intégral, *Histoire de l'Académie de Berlin* (1746), 182–224.
24. Davie, George Elder, *The Democratic Intellect: Scotland and her Universities in the Nineteenth Century*, The University Press, Edinburgh, 1966.
25. De Moivre, Abraham, *Miscellanea analytica*, Tonson and Watts, London, 1730.
26. Dijksterhuis, E. J., *Archimedes*, 2d. ed., Tr. C. Dikshoorn, Princeton University Press, Princeton, 1987.
27. Enneper, Alfred, *Elliptische Functionen: Theorie und Geschichte*, 2d. ed., Nebert, Halle, 1896.
28. Euler, Leonhard, De reductione formularum integralium ad rectificationem ellipsis ac hyperbolae, *Nov. Comm. Petrop.* 10 (1764), 30–50, in L. Euler, *Opera Omnia*, Teubner, Leipzig, Berlin, Zurich, 1911–, Ser. I, vol. 20, 256–301.
29. Euler, Leonhard, *Institutiones calculi differentialis*, 1755, sections 253–255. In *Opera*, Ser. I, Vol. XI.
30. Euler, Leonhard, *Introductio in analysin infinitorum*, Lausanne, 1748, in *Opera*, Ser. I, vols. 8–9.

31. Euler, Leonhard, Inventio summae cuiusque seriei ex data termino generali, *Comm. Petrop.* 8 (1741), 9–22; in *Opera*, Ser. I, vol. 14, 108–123.
32. Euler, Leonhard, Methodus generalis summandi progressiones, *Comentarii Acad. Imper. Petrop.* 6 (1738), 68–97; in *Opera*, Ser. II, vol. 22.
33. Euler, Leonhard, Methodus universalis serierum convergentium summas quam proxime inveniendi, *Comm. Petrop.* 8 (1741), 3–9, in *Opera*, Ser. I, vol. 14, 101–107.
34. Eyles, V. A., Hutton, *Dictionary of Scientific Biography*, vol. 6, 577–589.
35. Goldstine, Herman, *A History of Numerical Analysis from the 16th through the 19th Century*, Springer-Verlag, New York, Heidelberg, Berlin, 1977.
36. Grabiner, Judith V., A Mathematician Among the Molasses Barrels: MacLaurin's Unpublished Memoir on Volumes, *Proceedings of the Edinburgh Mathematical Society* 39 (1996), 193–240.
37. Grabiner, Judith V., *The Calculus as Algebra: J.-L. Lagrange, 1736–1813*, Garland Publishing, Boston, 1990.
38. Grabiner, Judith V., *The Origins of Cauchy's Rigorous Calculus*, M.I.T. Press, Cambridge, Mass., 1981.
39. Greenberg, John L., *The Problem of the Earth's Shape from Newton to Clairaut*, Cambridge University Press, Cambridge, 1995.
40. Griffiths, P. L., Private communication.
41. Griffiths, P. L., The British Influence on Euler's Early Mathematical Discoveries, preprint.
42. Guicciardini, Niccolo, *The Development of Newtonian Calculus in Britain, 1700–1800*, Cambridge University Press, Cambridge, 1989.
43. Hall, A. Rupert, *Philosophers at War: The Quarrel between Newton and Leibniz*, Cambridge University Press, Cambridge, 1980.
44. Hankins, Thomas, *Jean d'Alembert: Science and the Enlightenment*, Clarendon Press, Oxford, 1970.
45. Itard, Jean, Legendre, *Dictionary of Scientific Biography*, vol. 8, 135–143.
46. Jacobi, C. G. J., De usu legitimo formulae summatoriae Maclaurinianae, *Journ. f. reine u. angew. Math.* 18 (1834), 263–272. Also in *Gesammelte Werke*, vol. 6, 1891, pp. 64–75.
47. Juskevič, A. P., and R. Taton, eds., *Leonhard Euleri Commercium Epistolicum*, Birkhauser, Basel, 1980. In Leonhard Euler, *Opera*, Ser. 4, vol. 5.
48. Juskevič, A. P., and Winter, E., eds. *Leonhard Euler und Christian Goldbach: Briefwechsel, 1729–1764*, Akademie-Verlag, Berlin, 1965.
49. Kline, Morris, *Mathematical Thought from Ancient to Modern Times*, Oxford, New York, 1972.
50. Lacroix, S. F., *Traité élémentaire de calcul différentiel et de calcul intégral*, Duprat, Paris, 1802. Translated as **[51]**.
51. Lacroix, S.-F., *An Elementary Treatise on the Differential and Integral Calculus*, translated, with an Appendix and Notes, by C. Babbage, J. F. W. Herschel, and G. Peacock, J. Deighton and Sons, Cambridge, 1816.
52. Lacroix, S.-F., *Traité du calcul différentiel et du calcul intégral*, 3 vols., 2d. ed., Courcier, Paris, 1810–1819. Vol. I, 1810.
53. Lagrange, J.-L., Sur une nouvelle espèce de calcul rélatif a la différentiation et à l'intégration des quantités variables, *Nouvelles Memoires de l'académie... de Berlin*, 1772, 185–221, in *Oeuvres*, vol. 3, 439–476.
54. Lagrange, J.-L., *Leçons sur le calcul des fonctions*, new ed., Courcier, Paris, 1806. In *Oeuvres*, vol. 10.
55. Lagrange, J.-L., *Méchanique analytique*, 2d. ed., 2 vols., Courcier, Paris, 1811–1815, in *Oeuvres*, vols. 11–12.
56. Lagrange, J.-L., Note sur la métaphysique du calcul infinitésimal, *Miscellanea Taurinensia* 2 (1760–61), 17–18; in *Oeuvres*, vol. 7, 597–599.
57. Lagrange, J.-L., Sur l'attraction des sphéroïdes elliptiques, *Mémoires de l'académie de Berlin*, 1773, 121–148. Reprinted in *Oeuvres de Lagrange*, vol. III, 619ff.
58. Lagrange, Joseph-Louis, *Théorie des fonctions analytiques*, Imprimérie de la République, Paris, An V [1797]; compare the second edition, Courcier, Paris, 1813, reprinted in *Oeuvres de Lagrange*, pub. M. J.-A. Serret, 14 volumes, Gauthier-Villars, Paris, 1867–1892, reprinted again, Georg Oms Verlag, Hildesheim and New York, 1973, vol. 9.
59. Legendre, Adrien-Marie, Mémoires sur les intégrations par arcs d'ellipse et sur la comparaison de ces arcs, *Mémoires de l'Academie des sciences*, 1786, 616, 644–673.
60. Legendre, Adrien-Marie, *Traité des fonctions elliptiques et des intégrales eulériennes, avec des tables pour en faciliter le calcul numérique*, 3 vols., Paris, 1825–1828.
61. Loria, Gino, The Achievements of Great Britain in the Realm of Mathematics, *Mathematical Gazette* 8 (1915), 12–19.

62. Maclaurin, Colin, *Traité de fluxions*, Traduit de l'anglois par le R. P. Pézénas, 2 vols., Jombert, Paris, 1749.
63. Maclaurin, Colin, *A Treatise of Fluxions in Two Books*, Ruddimans, Edinburgh, 1742.
64. Mahoney, Michael, Review of [**42**], *Science* 250 (1990), 144.
65. McElroy, Davis, *Scotland's Age of Improvement*, Washington State University Press, Pullman, 1969.
66. Mills, Stella, *The Collected Letters of Colin Maclaurin*, Shiva Publishing, Nantwich, 1982.
67. Mills, Stella, The Independent Derivations by Leonhard Euler and Colin Maclaurin of the Euler-MacLaurin Summation Formula, *Archive for History of Exact Science* 33 (1985), 1–13.
68. Murdoch, Patrick, An Account of the Life and Writings of the Author, in Colin Maclaurin, *Account of Sir Isaac Newton's Philosophical Discoveries*, For the Author's Children, London, 1748, i-xx; reprinted, Johnson Reprint Corp., New York, 1968.
69. Newton, Isaac, *Of Analysis by Equations of an Infinite Number of Terms*, J. Stewart, London, 1745, in D. T. Whiteside, ed., *Mathematical Works of Isaac Newton*, vol. I, Johnson Reprint, New York and London, 1964, 3–25.
70. Olson, Richard, *Scottish Philosophy and British Physics, 1750–1880*, Princeton University Press, Princeton, 1975.
71. Olson, Richard, Scottish Philosophy and Mathematics, 1750–1830, *Journal of the History of Ideas* 32 (1971), 29–44.
72. Pearson, Karl, *The History of Statistics in the Seventeenth and Eighteenth Centuries* [written 1921–1933]. Ed. E. S. Pearson, Charles Griffin & Co., London and High Wycombe, 1976.
73. Pepe, Luigi, Tre 'prime edizioni' ed un' introduzione inedita della Fonctions analytiques di Lagrange, *Boll. Stor. Sci. Mat.* 6 (1986), 17–44.
74. Phillipson, Nicholas, The Scottish Enlightenment, in [**76**], pp. 19–40.
75. Porter, Roy, The Enlightenment in England, in [**76**], pp. 1–18.
76. Porter, Roy, and Mikulas Teich, eds., *The Enlightenment in National Context*, Cambridge University Press, Cambridge, 1981.
77. Playfair, John, Traité de Mechanique Celeste, *Edinburgh Review* 22 (1808), 249–84.
78. Sageng, Erik Lars, *Colin Maclaurin and the Foundations of the Method of Fluxions*, unpublished Ph.D. Dissertation, Princeton University, 1989.
79. Scott, J. F., Maclaurin, *Dictionary of Scientific Biography*, vol. 8, 609–612.
80. Shapin, Stephen, and Arnold Thackray, Prosopography as a Research Tool in History of Science: The British Scientific Community, 1700–1900, *History of Science* 12 (1974), 95–121.
81. Stewart, M. A., ed., *Studies in the Philosophy of the Scottish Enlightenment*, Clarendon Press, Oxford, 1990.
82. Struik, D. J., *A Source Book in Mathematics, 1200–1800*, Harvard University Press, Cambridge, Mass., 1969.
83. Taton, Juliette, Pezenas, *Dictionary of Scientific Biography*, vol. 10, Scribner's, New York, 1974, 571–2.
84. Taton, René, ed., *Enseignement et diffusion des sciences en France au XVIII^e siècle*, Hermann, Paris, 1964.
85. Todhunter, Isaac, *A History of the Mathematical Theories of Attraction and the Figure of the Earth, from the Time of Newton to That of Laplace*, Macmillan, London, 1873.
86. Truesdell, C., Rational Fluid Mechanics, 1687–1765, introduction to Euler *Opera*, Ser. 2, vol. 12.
87. Truesdell, C., The Rational Mechanics of Flexible or Elastic Bodies, 1638–1788, in Euler, *Opera*, Ser. 2, vol. 11.
88. Turnbull, H. W., *Bicentenary of the Death of Colin Maclaurin (1698–1746)*, The University Press, Aberdeen, 1951.
89. Turnbull, H. W., *The Great Mathematicians*, Methuen, London, 1929.
90. Tweedie, Charles, A Study of the Life and Writings of Colin Maclaurin, *Mathematical Gazette* 8 (1915), 132–151.
91. Tweedie, Charles, *James Stirling*, Clarendon Press, Oxford, 1922.

Pitzer College
1050 North Mills Avenue
Claremont, CA 91711, U.S.A.
jgrabiner@pitzer.edu

Yet Another Definition of Chaos

Pat Touhey

1. INTRODUCTION. The purpose of this article is to introduce yet another definition of chaos. The most popular and widely utilized definition is due to Devaney [**2**]. Namely, a self-map on a metric space X is chaotic on X if it has three essential ingredients: the periodic points of the map must form a dense subset of X, the map must have sensitive dependence on initial conditions, and the map must be topologically transitive.

In a paper appearing in this Monthly [**1**] Banks et al. showed that the hypothesis concerning sensitive dependence is implied by the remaining two conditions. Crannell points out in another recent Monthly article [**3**] that this mathematically elegant result yields a somewhat less intuitive definition of chaos. Of the three conditions that Devaney states, sensitive dependence is clearly the most easily understood. In order to restore this lost sense of intuitiveness, Crannell suggests a slightly more natural concept, *blending*, as an alternative to transitivity. While the idea of blending seems to be quite interesting in its own right, it is clearly demonstrated in [**3**] that blending and transitivity are not equivalent. We therefore propose a new definition of chaos, equivalent to Devaney's, and, we hope, just as natural. This definition arose several years ago in conversations with John Taylor about the article by Banks, et al. We reformulated the two topological conditions of transitivity and dense periodic points as a single condition that yields a simple, concise definition of chaos: A map $f: X \to X$ is chaotic on X if every pair of non-empty open subsets of X shares a periodic orbit. We use our definition to give a characterization of chaos that restores the lost sense of intuitiveness: A map $f: X \to X$ is chaotic on X if and only if it mixes together, via periodic cycles, any finite number of non-empty open subsets in infinitely many ways.

2. CHAOS

Definition 2.1. *Given a non-empty set X and a mapping $f: X \to X$, the forward orbit of x under f is the set $O_f^+(x) \equiv \{x, f(x), f^2(x), \dots\}$, where $f^n(x) = f(f^{n-1}(x))$ for $n \geq 1$ with $f^0(x) \equiv x$.*

Definition 2.2. *Given a non-empty set X and a mapping $f: X \to X$, x is a periodic point of f with primitive period n if $f^n(x) = x$ but $f^k(x) \neq x$ for $k = 1, \dots, n-1$.*

We now define the concept of topological transitivity.

Definition 2.3. *Given a metric space X and a continuous mapping $f: X \to X$, we say that f is transitive if for any two non-empty open subsets U and V of X, there exists some $u \in U$ and a non-negative integer k such that $f^k(u) \in V$, that is, every pair of non-empty open subsets of X shares a forward orbit.*

Although this is not the usual definition of transitivity, we have no doubt the reader can prove the following assertion of equivalence between it and the more common notion.

Proposition 2.4. *$f: X \to X$ is transitive if and only if for any two non-empty open sets $U, V \subset X$ there exists a non-negative integer k such that $f^k(U) \cap V \neq \varnothing$.*

And now we turn to the major result of this article: Yet another definition of chaos.

Definition 2.5. *Given a metric space X and a continuous mapping $f: X \to X$, we say that f is chaotic on X if given U and V, non-empty open subsets of X, there exists a periodic point $p \in U$ and a non-negative integer k such that $f^k(p) \in V$, that is, every pair of non-empty open subsets of X shares a periodic orbit.*

It now remains to show that our new definition of chaos is equivalent to the definition given by Devaney in [**2**]. By our remarks in the first section, it is sufficient to show that our definition is equivalent to the pair of conditions that f be transitive and have dense periodic points.

Proposition 2.6. *$f: X \to X$ is chaotic on X if and only if f is transitive and the periodic points of f are dense in X.*

Proof: If f is chaotic on X then every pair of non-empty open sets shares a periodic orbit. In particular, every non-empty open set must contain a periodic point so the periodic points of f are dense in X. The transitivity of f follows from the definition of chaos since every pair of non-empty open sets shares a forward orbit.

Now let us assume that f is transitive and has a dense set of periodic points. Given any pair of non-empty open sets $U, V \subset X$ transitivity ensures that there exists $u \in U$ and a non-negative integer k such that $f^k(u) \in V$. Now define $W \equiv f^{-k}(V) \cap U$. Note that W is open and non-empty since it is the intersection of two open sets and u is an element of both of them. It is also clear that W has the property that $f^k(W) \subset V$. But the periodic points of f are assumed to be dense in X, so the non-empty open set W must contain a periodic point p. Thus we have shown that there exists a periodic point $p \in W \subset U$ with the property that $f^k(p) \in f^k(W) \subset V$. This implies that f is chaotic. ■

It can now be shown that any finite number of non-empty open sets shares a periodic orbit whenever f is chaotic.

Proposition 2.7. *Let X be a metric space and let $f: X \to X$ be a chaotic mapping on X. Then any finite collection of non-empty open subsets of X shares a periodic orbit.*

Proof: Let N be the number of non-empty open subsets in our collection. If $N = 1$, the result follows from the density of periodic points; if $N = 2$ it follows from the definition of a chaotic mapping. We proceed by induction on N. Thus assume that the assertion holds for $N = n$. We will show that it holds for $n + 1$ non-empty open subsets.

There is no loss of generality to assume that the collection consists of $n + 1$ disjoint subsets. If the sets are not disjoint then some pair of the non-empty open subsets intersects in an open subset. Replacing the pair by their intersection yields a collection of n non-empty open subsets that, by our induction hypothesis, shares a periodic orbit. Clearly this orbit is shared by the original collection of $n + 1$ subsets.

Now from our disjoint collection choose a subset and call it V. The remaining n subsets must share a periodic orbit and, like all periodic orbits, this orbit has a

primitive period, which we designate by M. From these remaining n non-empty open subsets choose any subset and call it U_0. Thus we must have $p \in U_0$, where p is a periodic point of primitive period $M > n - 1$, with the property that $O_f^+(p)$ intersects each of our n subsets. We now label each of the remaining $n - 1$ non-empty open subsets in the following manner. As we iterate the point p, it first intersects one of the $n - 1$ open subsets for some value, k_1, of the iterate, $0 < k_1 < M$. Let this subset be designated by U_1, i.e., $f^{k_1}(p) \in U_1$. Continuing in this fashion we arrive at the next iterate, $f^{k_2}(p)$, $0 < k_1 < k_2 < M$, intersecting one of the remaining $n - 2$ open subsets. This subset is designated U_2. Eventually we will have labeled each of the n open subsets so that $f^{k_i}(p) \in U_i$ for all $i = 0, 1, \ldots, n - 1$, where $0 = k_0 < k_1 < \cdots < k_{n-1} < M$. Now we define another collection of non-empty open subsets with a particularly nice property. Let $W_0 \equiv U_{n-1}$. Clearly $f^{k_{n-1}}(p) \in W_0$. Now consider

$$W_1 \equiv f^{-[k_{n-1}-k_{n-2}]}(W_0) \cap U_{n-2}.$$

We claim that W_1 is a non-empty open subset contained in U_{n-2}. It is open because it is the intersection of two open subsets, and it is obviously in U_{n-2}. That it is non-empty follows from the facts that $f^{k_{n-1}}(p) \in W_0$ and $f^{k_{n-2}}(p) \in U_{n-2}$, and hence

$$f^{k_{n-2}}(p) = f^{-[k_{n-1}-k_{n-2}]}(f^{k_{n-1}}(p)) \in f^{-[k_{n-1}-k_{n-2}]}(W_0),$$

which implies that $f^{k_{n-2}}(p) \in W_1$. Also note that W_1 has the particularly nice property that $f^{[k_{n-1}-k_{n-2}]}(W_1) \subset W_0$. Continuing in this fashion, we define

$$W_i \equiv f^{-[k_{n-i}-k_{n-(i+1)}]}(W_{i-1}) \cap U_{n-(i+1)} \quad \text{for } i = 1, 2, \ldots, n - 1.$$

Each W_i is again non-empty, open, and contained in $U_{n-(i+1)}$. In addition we have the particularly nice property that

$$f^{[k_{n-i}-k_{n-(i+1)}]}(W_i) \subset W_{i-1} \quad \text{for } i = 1, 2, \ldots, n - 1.$$

It is easy to find a periodic orbit that wends itself through our original collection of $n + 1$ non-empty open subsets, $\{V, U_0, U_1, \ldots, U_{n-1}\}$. Since V and W_{n-1} are both open, they share a periodic orbit. Thus, there exists a periodic point $p' \in V$ and a positive integer q such that $f^q(p') \in W_{n-1} \subset U_0$. But then, by our particularly nice property, the subsequent iterates of p' must pass through all of the U_i's.

$$\begin{aligned}
f^q(p') &= f^{[q+k_0]}(p') \in W_{n-1} \subset U_0\\
f^{q+k_1}(p') &= f^{[k_1-k_0]}(f^{[q+k_0]}(p')) \in f^{[k_1-k_0]}(W_{n-1}) \subset W_{n-2} \subset U_1\\
&\vdots\\
f^{q+k_i}(p') &= f^{[k_i-k_{i-1}]}(f^{[q+k_{i-1}]}(p')) \in f^{[k_i-k_{i-1}]}(W_{n-i}) \subset W_{n-(i+1)} \subset U_i\\
&\vdots\\
f^{q+k_{n-1}}(p') &= f^{[k_{n-1}-k_{n-2}]}(f^{[q+k_{n-2}]}(p')) \in f^{[k_{n-1}-k_{n-2}]}(W_1) \subset W_0 = U_{n-1}.
\end{aligned}$$

Thus, the forward orbit of p', $O_f^+(p')$, intersects each of $V, U_0, U_1, \ldots, U_{n-1}$. ■

Corollary 2.8. *Let X be a metric space and let $f\colon X \to X$ be a chaotic mapping on X. Then any finite collection of non-empty open subsets of X shares infinitely many periodic orbits.*

Proof: Assume the existence of a finite collection $\{U_i\}_{i=1,\dots,n}$ of non-empty open subsets that share only a finite number of periodic orbits. Define P to be the set consisting of the union of the points in these shared periodic orbits. Since each periodic orbit contains a finite number of points, the union of finitely many such orbits must be finite. Hence P is a finite set. We now define another collection of non-empty open subsets $\{V_i\}_{i=1,\dots,n}$ by $V_i \equiv U_i \setminus P$. It's clear that each $V_i \subset U_i$. And each V_i is non-empty and open since removing the finite set of points, P, from the open set U_i leaves us with a non-empty open set. Thus by Proposition 2.7 there must be a periodic orbit shared by the collection $\{V_i\}_{i=1,\dots,n}$. This new orbit is clearly not contained in P. On the other hand, this orbit obviously passes through the original collection $\{U_i\}_{i=1,\dots,n}$ of non-empty open subsets since each $V_i \subset U_i$. This contradiction proves our result. ■

Proposition 2.9. *Let X be a metric space and f: $X \to X$ be a mapping. The following are equivalent:*

i) *f is chaotic on X.*
ii) *f is topologically transitive and has a dense set of periodic points.*
iii) *any finite collection of non-empty open sets of X shares a periodic orbit.*
iv) *any finite collection of non-empty open sets of X shares infinitely many periodic orbits.*

Proof: We have already shown that i) $\Leftrightarrow$ ii) $\Rightarrow$ iii) $\Rightarrow$ iv).

If any finite collection of non-empty open sets contained in X shares infinitely many periodic orbits it is clear that any pair of open sets shares a periodic orbit. Thus iv) $\Rightarrow$ i). ■

REFERENCES

1. J. Banks, J. Brooks, G. Cairns, G. Davis, and P. Stacey, On Devaney's Definition of Chaos, *Amer. Math. Monthly* **99** (1992) 332–334.
2. R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, Addison-Wesley, Redwood City, Calif., 1989.
3. A. Crannell, The Role of Transitivity in Devaney's Definition of Chaos, *Amer. Math. Monthly* **102** (1995) 788–793.

Department of Mathematics and Computer Science
College Misericordia
Dallas, Pennsylvania 18612
Touhey@aol.com

...I drove up the mountain and found a dairy, bought some milk, and asked permission to camp under an apple tree. The dairy man had a Ph.D. in mathematics, and he must have had some training in philosophy. He liked what he was doing and he didn't want to be somewhere else—one of the very few contented people I met in my whole journey.

John Steinbeck, *Travels with Charley*
Viking Press, New York, 1962, pp. 25–26
Contributed by Harold P. Boas, Texas A & M University

Hereditary Classes of Operators and Matrices

Scott A. McCullough and Leiba Rodman

1. INTRODUCTION. A recent problem in this Monthly [**19**] asks whether the identity

$$A^*A = A^2 \tag{1.1}$$

implies that the matrix A is hermitian. Here A^* denotes the conjugate transpose of the matrix A. The answer is yes (of course, only the case of singular A is nontrivial), as can be proved by elementary methods: By Schur's triangularization theorem, which asserts that every square size complex matrix is unitarily similar to an upper triangular matrix [**17**, Theorem 2.3.1], we may assume that A itself is upper triangular. Inspection of the diagonal entries of $A^*A = A^2$ shows that all the diagonal entries of A are real and all the off-diagonal entries are zero. Thus, A is Hermitian, since it is unitarily similar to a real diagonal matrix.

Consider now a "symmetrized" version of (1.1) in which the squares of both A and A^* appear symmetrically:

$$A^2 - 2A^*A + A^{*2} = 0. \tag{1.2}$$

Clearly, any matrix A that satisfies (1.1) also satisfies (1.2). Does (1.2) imply that A is hermitian? The answer is again yes, and, moreover, it is yes also for a bounded linear operator acting on an infinite-dimensional Hilbert space (Theorem 3.1). Thus, in the algebra of bounded linear operators on a fixed Hilbert space, the three operator identities (1.1), (1.2), and $A = A^*$ are equivalent.

This result is a very special case of the theory of hereditary classes of bounded linear operators developed in [**2**]. The purpose of this paper is to expose the general framework of hereditary classes, including the main ideas, results, several particular cases, examples, open problems, and applications; this will be done in Section 4. In the next two sections we study certain hereditary classes of particular interest (n-selfadjoint operators) in detail. Here we mention only that many hereditary classes are defined in terms of operator identities of the form

$$p(A^*, A) = 0, \tag{1.3}$$

where $p(\lambda, \mu)$ is a polynomial with complex coefficients in two non-commuting variables λ and μ such that in every summand of $p(\lambda, \mu)$ the powers of λ (if any) appear on the left of the powers of μ (if any). For example, the class of all bounded selfadjoint operators is a hereditary class associated with the polynomial $p(\lambda, \mu) = \lambda - \mu$.

In a broad and admittedly imprecise sense, the theory of hereditary classes can be thought of as a "noncommutative" (namely, an operator and its adjoint do not necessarily commute) spectral theory, which generalizes and extends the well-known "commutative" spectral theory for bounded selfadjoint, or, more generally, normal, operators.

Section 5 is devoted mainly to several classes of operators that are close in a certain sense to selfadjoint operators, in spaces with an indefinite inner product. The results there are a modest beginning of a theory of hereditary classes in the context of indefinite inner product spaces. Since this is essentially uncharted territory, there are many open problems, and we state a few of them.

2. n-SELFADJOINT OPERATORS: THE FINITE DIMENSIONAL CASE. To start with, we consider a natural generalization of the formula (1.2). A bounded linear operator A on a Hilbert space H is called *n-selfadjoint* if

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} A^{*k} A^{n-k} = 0, \tag{2.1}$$

where n is a fixed positive integer; A^0 and A^{*0} are interpreted as the identity operator I. Obviously, 1-selfadjoint operators are just selfadjoint. We immediately make the following simple observation that will be used later on:

Proposition 2.1. *If A is n-selfadjoint, then for every real number μ the operator $A - \mu I$ is n-selfadjoint as well.*

Proof: Write

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} (A - \mu I)^{*k} (A - \mu I)^{n-k}$$
$$= \sum_{k=0}^{n} (-1)^k \binom{n}{k} A^{*k} A^{n-k} + \sum_{p+q<n} A^{*p} A^q f(p, q, \mu),$$

where $f(p, q, \mu)$ is a certain number. We have

$$f(p, q, \mu) = \sum (-1)^k \binom{n}{k} \binom{k}{p} (-\mu)^{k-p} \binom{n-k}{q} (-\mu)^{n-k-q},$$

where the summation is over all integers k such that $p \le k \le n - q$. An easy calculation shows that

$$f(p, q, \mu) = (-\mu)^{n-p-q} \sum_{k=p}^{n-q} (-1)^k \binom{n}{k} \binom{k}{p} \binom{n-k}{q}$$
$$= (-\mu)^{n-p-q} \frac{n!}{(n-p-q)!p!q!} \sum_{k=p}^{n-q} (-1)^k \binom{n-q-p}{k-p} = 0. \quad \blacksquare$$

It turns out that the class of n-selfadjoint operators is hereditary with

$$p(\lambda, \mu) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \lambda^k \mu^{n-k}.$$

The structure theorem for n-selfadjoint matrices asserts the following:

Theorem 2.2. *An operator A on a finite dimensional Hilbert space is n-selfadjoint if and only if*

$$A = T + N, \tag{2.2}$$

where T is selfadjoint, $N^{[\frac{n+1}{2}]} = 0$, and T and N commute. ([r] *denotes the integer part of r*).

An interesting feature of this result is that for any integer $k > 0$, the $2k$-selfadjoint matrices and the $(2k-1)$-selfadjoint matrices are exactly the same class. As we will see, the same is true for operators. We verify in Section 3 that every operator of the form (2.2), in which T and N have the properties indicated in Theorem 2.2, is n-selfadjoint, even when the underlying Hilbert space is infinite dimensional. It turns out that the converse is nearly true (at least, for the 3-selfadjoint operators). This is also discussed in Section 3.

We now prove the difficult direction "only if" of Theorem 2.2, using elementary linear algebra. Thus, we assume throughout the rest of this section that $H = \mathbf{C}^p$, the vector space of columns with p complex components, and that A is a $p \times p$ matrix that satisfies (2.1) (for a fixed n).

Lemma 2.3. *If x and y are eigenvectors of A corresponding to the eigenvalues λ and μ, respectively, then*

$$(\lambda - \bar{\mu})y^*x = 0. \tag{2.3}$$

Proof: We have $Ax = \lambda x$ and $y^*A^* = \bar{\mu}y^*$. Thus

$$0 = y^*\left(\sum_{k=0}^{n}(-1)^k\binom{n}{k}A^{*k}A^{n-k}\right)x$$
$$= \sum_{k=0}^{n}(-1)^k\binom{n}{k}\bar{\mu}^k(y^*x)\lambda^{n-k} = (\lambda - \bar{\mu})^n y^*x,$$

and (2.3) follows. ■

The next observation is an immediate consequence of Lemma 2.3.

Lemma 2.4. *All eigenvalues of A are real.*

Indeed, if λ were a non-real eigenvalue of A, then taking $\mu = \lambda$, $y = x$ in (2.3) would lead to a contradiction.

Lemma 2.5. *If λ and μ are distinct eigenvalues of A, then the root subspaces* $\operatorname{Ker}(A - \lambda I)^p$ *and* $\operatorname{Ker}(A - \mu I)^p$ *are orthogonal.*

Proof: Let y be an eigenvector of A corresponding to the eigenvalue μ (necessarily real, by Lemma 2.4). Then for every $x \in \mathbf{C}^p$ we have

$$0 = y^*\left(\sum_{k=0}^{n}(-1)^k\binom{n}{k}A^{*k}A^{n-k}\right)x$$
$$= y^*\left(\sum_{k=0}^{n}(-1)^k\binom{n}{k}\mu^kA^{n-k}\right)x = y^*(\mu I - A)^n x.$$

Thus, $y \perp \operatorname{Im}(\mu I - A)^n$. The nonincreasing sequence of subspaces

$$\operatorname{Im}(\mu I - A) \supseteq \operatorname{Im}(\mu I - A)^2 \supseteq \cdots$$

must stabilize because of the finite dimensionality of $\mathbf{C}^p$. Let q be a positive integer such that

$$\operatorname{Im}(\mu I - A)^q = \operatorname{Im}(\mu I - A)^r$$

for all $r > q$. We certainly have $y \perp \operatorname{Im}(\mu I - A)^q$. If we have proved already that $(A - \mu I)^j u \perp \operatorname{Im}(\mu I - A)^q$ for $j = 0, 1, \ldots,$ and if $v \in \mathbf{C}^p$ is such that

$(A - \mu I)v = u$, then for any $y \in \operatorname{Im}(\mu I - A)^q$, Proposition 2.1 ensures that

$$0 = v^*\left(\sum_{k=0}^{n}(-1)^k\binom{n}{k}(A^* - \mu I)^k(A - \mu I)^{n-k}\right)y = v^*(A - \mu I)^n\, y.$$

Thus, $v \perp \operatorname{Im}(\mu I - A)^{n+q} = \operatorname{Im}(\mu I - A)^q$. Starting with the eigenvectors of A corresponding to μ, induction and the preceding argument show that $\operatorname{Ker}(\mu I - A)^p \perp \operatorname{Im}(\mu I - A)^q$. Since $\operatorname{Ker}(A - \lambda I)^p \subseteq \operatorname{Im}(\mu I - A)^q$,

$$\operatorname{Ker}(A - \lambda I)^p \perp \operatorname{Ker}(A - \mu I)^p$$

follows. ∎

In view of Lemma 2.5, each root subspace $\operatorname{Ker}(A - \lambda I)^p$ is A^*-invariant (as well as A-invariant). Thus, we need to prove Theorem 2.2 for the restriction of A to each root subspace $\operatorname{Ker}(A - \lambda I)^p$ separately. In other words, we may assume that A has only one eigenvalue (possibly of high multiplicity), and in view of Proposition 2.1 we may further assume that this eigenvalue is 0, i.e., A is nilpotent. Thus, the proof of the "only if" part of Theorem 2.2 reduces to the following lemma.

Lemma 2.6. *If A is nilpotent and satisfies* (2.1), *then* $A^{[\frac{n+1}{2}]} = 0$.

Proof: Arguing by contradiction, it follows easily from the Jordan form of A that there exists a vector x in $\mathbf{C}^p$ such that $A^{[\frac{n+1}{2}]}x = y \neq 0$ and $A^q x = 0$ for all $q > \left[\frac{n+1}{2}\right]$. Assume first that n is even, say $n = 2m$. Then

$$\begin{aligned} 0 &= x^*\left(\sum_{k=0}^{n}(-1)^k\binom{n}{k}A^{*k}A^{n-k}\right)x \\ &= \sum_{k=0}^{n}(-1)^k\binom{n}{k}x^*A^{*k}A^{n-k}x = (-1)^m\binom{2m}{m}y^*y, \end{aligned}$$

a contradiction. If n is odd, say, $n = 2m + 1$, then the preceding formula has the analog

$$\begin{aligned} 0 &= (Ax)^*\left(\sum_{k=0}^{n}(-1)^k\binom{n}{k}A^{*k}A^{n-k}\right)x \\ &= \sum_{k=0}^{n}(-1)^k\binom{n}{k}x^*A^{*k+1}A^{n-k}x = (-1)^m\binom{2m+1}{m}y^*y, \end{aligned}$$

a contradiction with $y \neq 0$. ∎

For the special case of 3-selfadjoint matrices, Theorem 2.2 can be recast in terms of a canonical form under unitary similarity (two matrices X and Y are called *unitarily similar* if $X = UYU^*$ for some unitary matrix U). This result (Theorem 2.7) is in the spirit of the spectral theory and illustrates how Theorem 2.2 can be thought of as an extension of the familiar spectral theorem for Hermitian matrices. An operator X on a finite dimensional Hilbert space is called *indecomposable* if there is no nontrivial subspace that is invariant for both X and X^*.

Theorem 2.7. *The* 1×1 *matrices* $[c]$ *and* 2×2 *matrices* $\begin{bmatrix} c & d \\ 0 & c \end{bmatrix}$ (*where c is real and* $d > 0$) *are the only* (*up to unitary similarity*) *indecomposables in the class of* 3-*selfadjoint matrices. Moreover, every* 3-*selfadjoint matrix* A *is unitarily similar to an orthogonal sum of indecomposables; this orthogonal sum is uniquely determined, up to permutations of the orthogonal summands, by* A.

This follows easily from Theorem 2.2 using the following fact: *Every* $p \times p$ *matrix* N *such that* $N^2 = 0$ *is unitarily similar to a unique* (*up to a permutation of the orthogonal summands*) *orthogonal sum*

$$\begin{bmatrix} 0 & d_1 \\ 0 & 0 \end{bmatrix} \oplus \cdots \oplus \begin{bmatrix} 0 & d_s \\ 0 & 0 \end{bmatrix} \oplus [0] \oplus \cdots \oplus [0], \tag{2.4}$$

where $d_1, \dots, d_s$ *are positive numbers*. The decomposition (2.4) follows from representing N as a block matrix

$$\begin{bmatrix} 0 & 0 & N_0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

with respect to the orthogonal decomposition

$$\mathbf{C}^p = \operatorname{Im} N \oplus (\operatorname{Ker} N \ominus \operatorname{Im} N) \oplus (\operatorname{Ker} N)^{\perp},$$

and a subsequent reduction of N_0 to a positive diagonal matrix by unitary equivalence ($N_0 \to UN_0V$, where U and V are unitary matrices; this is the *singular value decomposition* of N_0). The uniqueness of (2.4) follows from observing that $s = \operatorname{rank} N$, and $d_1, \dots, d_s$ are the nonzero singular values of N.

3. n**-SELFADJOINT OPERATORS: THE INFINITE DIMENSIONAL CASE.** The development of the theory of n-selfadjoint operators in infinite dimensional Hilbert spaces was motivated largely by striking and unexpected intimate connections with differential equations, particularly conjugate point theory and disconjugacy. These connections were first observed by Helton **[16]**, and further exploited in **[3]**. Helton **[15, 16]** introduced the 3-selfadjoint operators (or, more precisely, 3-selfadjoint operators having a cyclic vector) as models for the "multiplication by x" operators on weighted Sobolev spaces supported on a compact interval $[a, b]$ (see the example given after Theorem 3.3).

The theory of n-selfadjoint operators acting on infinite dimensional Hilbert space differs significantly from that of n-selfadjoint matrices. In particular, Theorem 2.2 fails in general in infinite dimensions. In this section we discuss some aspects of the available results concerning infinite dimensional n-selfadjoint operators. Throughout the rest of the paper, we'll use $\mathscr{L}(H)$ to denote the bounded linear operators on the Hilbert space H and by operator we'll mean an element of $\mathscr{L}(H)$ for some H.

Theorem 3.1. *An operator* $A \in \mathscr{L}(H)$ *is* 2-*selfadjoint if and only if* A *is selfadjoint.*

In other words, the result of Theorem 2.2 holds for infinite dimensional operators if $n = 2$. The authors of **[2, 15, 8]** were well aware of Theorem 3.1, although none of these papers state it explicitly. We will see that Theorem 3.1 is a byproduct of more general results on operators in spaces with an indefinite scalar product (specifically, Theorem 5.1).

Let's agree to call an operator $J \in \mathscr{L}(H)$ a *Jordan operator* if $J = T + N$, where T is selfadjoint, $N^2 = 0$, and $TN = NT$. If $J \in \mathscr{L}(H)$ is a Jordan operator, then J is a 3-selfadjoint operator. We say that an operator A is *n-Jordan* if A has the form (2.2).

Theorem 3.2. *Every n-Jordan operator is n-selfadjoint.*

Proof: Suppose $J = T + N$ with T selfadjoint, $N^n = 0$, and $TN = NT$. We shall show that J is $2n + 1$-selfadjoint. We have

$$J^m = \sum_{j=0}^{n} \binom{m}{j} N^j T^{m-j}. \tag{3.1}$$

Using (3.1) and the properties of T, N, we obtain

$$\sum_{m=0}^{2n+1} (-1)^m \binom{2n+1}{m} J^{*k} J^{2n+1-m}$$

$$= \sum_{j,l=0}^{n} N^{*j} N^l T^{2n+1-j-l} \sum_{m=l}^{2n+1-j} (-1)^m \binom{2n+1}{m} \binom{m}{l} \binom{2n+1-m}{j}.$$

The summation on m simplifies to

$$\frac{(2n+1)!}{l!j!} \sum_{m=l}^{2n+1-j} (-1)^m \frac{1}{(2n+1-m-j)!(m-l)!}.$$

This last summation is of the form $\pm\Sigma_{q=0}^{p}(-1)^q(q!(p-q)!)^{-1}$ which is zero. We conclude that J is $2n + 1$-selfadjoint. ∎

If J is a Jordan operator and $M \subset H$ is an invariant subspace for J, then $A = J|_M$ is also a 3-selfadjoint operator. If H is finite dimensional, then A is again a Jordan operator, but in general (i.e., if H is infinite dimensional) A need not be a Jordan operator. The following theorem clarifies the relation between Jordan and 3-selfadjoint operators. An initial version of this theorem is due to Helton **[15]**. Agler **[1]** established the general result.

Theorem 3.3 (3-selfadjoint lifting). *If A is a 3-selfadjoint operator on a Hilbert space H, then there exists a Jordan operator J on a Hilbert space K containing H such that H is invariant for J and $A = J|_H$.*

In other words, A lifts to J (see the next section for a more detailed discussion of this notion). If, however, both A and A^* are 3-selfadjoint, then A is Jordan (a result proved in **[15]**).

The following example illustrates Theorem 3.3. Let H denote the Hilbert space obtained by taking the closure of the continuously differentiable functions on [0, 1] in the norm induced by the inner product

$$\langle p, q \rangle = \int_0^1 (p'\overline{q'} + p\overline{q})\,dt.$$

It is an entertaining exercise to verify directly that the operator A of multiplication by t on H, $Af(t) = tf(t)$, is bounded and 3-selfadjoint, but this fact will follow

shortly. Let Y denote the operator of multiplication by t on $L^2\,[0,1]$ and define

$$J = \begin{bmatrix} Y & I \\ 0 & Y \end{bmatrix},$$

acting on $K = L^2\,[0,1] \oplus L^2\,[0,1]$, where I denotes the identity operator. Define $V\colon H \mapsto K$ on continuously differentiable functions (so V is densely defined) by

$$Vp = \begin{bmatrix} p' \\ p \end{bmatrix}.$$

It is easy to see that V is an isometry and thus extends to an isometry, also denoted by V, on H. Further, $VA = JV$, so that VH is invariant for J, and, identifying VH with H, $A = J|_H$. Since J is bounded, so is A.

This last example also shows that not every 3-selfadjoint operator is Jordan. To see this begin by noting that if $J = T + N$ is Jordan, then $J^* = T + N^*$ is also Jordan and therefore 3-selfadjoint. The norm on H dominates the supremum norm, and hence for each $0 \le x \le 1$, there is some $k_x \in H$ such that $\langle p, k_x \rangle = p(x)$, for each polynomial p. (One could also show directly the existence of k_x; in particular, the formulas for k_0 and k_1 are given below). A routine argument shows $A^* k_x = x k_x$. We can thus compute

$$\langle (A^{*3} - 3AA^{*2} + 3A^2A^* - A^3)k_x, k_y \rangle = (x-y)^3 \langle k_x, k_y \rangle.$$

Hence, if A^* is 3-selfadjoint, then $\langle k_x, k_y \rangle = 0$, as long as $x \neq y$. But notice that $k_0 = (e-1)^{-1}(\exp(t) + \exp(2)\exp(-t))$ and $k_1 = (\cosh(1))^{-1}\cosh(t)$; these formulas for k_0 and k_1 can be verified using integration by parts to show that $\langle f, k_0 \rangle = f(0)$ and $\langle f, k_1 \rangle = f(1)$ for any continuously differentiable function f on $[0,1]$. However, $\langle k_0, k_1 \rangle \neq 0$.

Another interesting property of n-selfadjoint operators is the following:

Theorem 3.4. *For every positive integer k, the classes of $2k$-selfadjoint operators and $(2k-1)$-selfadjoint operators coincide.*

It is possible to give a proof of Theorem 3.4 based upon the techniques of **[15]**. Note that Theorem 3.1 is a particular case of Theorem 3.4.

4. HEREDITARY CLASSES AND FAMILIES OF OPERATORS. In this section we present an informal introduction to a part of the work of J. Agler on families of operators and hereditary polynomials, borrowing freely from **[2]**, and focusing on the relation between a family and its boundary. We'll illustrate the basics of the abstract theory with a family of 3-selfadjoint operators and a local family of contraction operators as examples. This local family of contractions is essentially finite dimensional. For the reader who prefers to think in terms of matrices, this provides an introduction both to families of operators and to a class of operators that is preeminent in operator theory. We'll first discuss hereditary polynomials and then introduce the definition of a family.

An *hereditary polynomial* is a polynomial in two noncommuting variables x, y of the form

$$p(x,y) = \sum c_{ij} y^j x^i.$$

Given an operator A, define

$$p(A) = \sum c_{ij} A^{*j} A^i.$$

If $q(x, y)$ is any polynomial then there is a unique hereditary polynomial $p(x, y)$ such that $p(z, w) = q(z, w)$ for all complex numbers z, w. In this case, define

$$q(A) = p(A).$$

For instance, if $q(x, y) = (x - y)^3$, then

$$q(A) = A^{*3} - 3A^{*2}A + 3A^{*}A^2 - A^3.$$

More generally, if $q(x, y) = (x - y)^n$, then

$$q(A) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} A^{*k}A^{n-k}.$$

Thus, A is an n-selfadjoint operator if and only if $(x - y)^n(A) = 0$.

An operator X on a Hilbert space H is positive, denoted $X \geq 0$, if $\langle Xh, h\rangle \geq 0$ for every $h \in H$. Thus, A is n-selfadjoint if and only if

$$\pm(x - y)^n(A) \geq 0.$$

Given $A \in \mathscr{L}(H)$ a subspace M of H (all subspaces are closed) is *invariant* for A provided $AM \subset M$. The restriction of A to M is denoted $A|_M$. The following lemma expresses the distinguishing property of hereditary polynomials.

Lemma 4.1. *Let A be an operator on a Hilbert space H and suppose M is invariant for A. If p is an hereditary polynomial and $p(A) \geq 0$, then $p(A|_M) \geq 0$.*

This lemma is a consequence of the fact that, for $h, k \in M$ and p an hereditary polynomial, $\langle p(A|_M)h, k\rangle = \langle p(A)h, k\rangle$. Here are the details with $h = k$. Writing $p(x, y) = \sum c_{ij}y^jx^i$ and given $h \in M$, we have

$$\begin{aligned}\langle p(A|_M)h, h\rangle &= \sum c_{ij}\langle (A|_M)^{*j}(A|_M)^i h, h\rangle \\ &= \sum c_{ij}\langle (A|_M)^i h, (A|_M)^j h\rangle \\ &= \sum c_{ij}\langle A^i h, A^j h\rangle \\ &= \sum c_{ij}\langle A^{*j}A^i h, h\rangle \\ &= \langle p(T)h, h\rangle,\end{aligned}$$

where in the third equality we use the fact that M is invariant for A.

Given a collection γ of hereditary polynomials, let $\mathscr{F}_\gamma$ denote the collection of operators $A \in \mathscr{L}(H)$ such that $p(A) \geq 0$ for every $p \in \gamma$. For reasons that will become clear, we will always assume that there is a positive constant c such that $c^2 - yx \in \gamma$.

Lemma 4.2. *If γ is a collection of hereditary polynomials and there exists a positive constant c such that $c^2 - yx \in \gamma$, then $\mathscr{F}_\gamma$ satisfies*

(i) *$\mathscr{F}_\gamma$ is closed with respect to orthogonal direct sums; i.e., if $A_\alpha \in \mathscr{F}_\gamma$, then $\oplus_\alpha A_\alpha \in \mathscr{F}_\gamma$;*
(ii) *$\mathscr{F}_\gamma$ is closed with respect to unital $*$-representations; i.e., if $A \in \mathscr{F}_\gamma \cap \mathscr{L}(H)$ and π is a $*$-homomorphism mapping I into I from the norm closed subalgebra generated by A, A^* and I into $\mathscr{L}(K)$, then $\pi(A) \in \mathscr{F}_\gamma$;*
(iii) *$\mathscr{F}_\gamma$ is closed with respect to restriction to invariant subspaces; i.e., if $A \in \mathscr{F}_\gamma \cap \mathscr{L}(H)$ and $M \subset H$ is invariant for A, then $A|_H \in \mathscr{F}_\gamma$;*
(iv) *$\mathscr{F}_\gamma$ is bounded. In fact, if $A \in \mathscr{F}_\gamma$, then the operator norm of A does not exceed c.*

The condition (iii) is a consequence of Lemma 4.1. The hypothesis $c^2 - yx \in \gamma$ directly implies (iv). The conditions (i) and (ii) are easily checked and, while important, will not be emphasized here. We say, as shorthand for the conclusion of Lemma 4.2, that $\mathscr{F}_\gamma$ is a bounded collection of operators that is closed with respect to orthogonal direct sums, unital $*$-representations, and restrictions to invariant subspaces. Any bounded collection of operators $\mathscr{F}$ that is closed with respect to orthogonal direct sums, unital $*$-representations, and restriction to invariant subspaces is called a *family*. While we haven't defined matrix valued hereditary polynomials, a theorem of Agler [2] says if $\mathscr{F}$ is a family, then there is a collection of matrix valued hereditary polynomials Γ such that $\mathscr{F} = \mathscr{F}_\Gamma$.

We now introduce a family of operators that serves as an example throughout the remainder of this section. An operator C on a Hilbert space H is a *contraction* if its operator norm is at most one. Equivalently, C is a contraction if

$$(1 - yx)(C) = I - C^*C \geq 0.$$

The collection of all contractions forms a family of operators that we denote by $\mathscr{C}$. We can obtain a family of contraction operators whose members behave like matrices by, in Agler's terminology, *localization*. To keep things simple, fix distinct points $0 = \lambda_0, \lambda_1, \dots, \lambda_n$ with $|\lambda_j| < 1$ and let

$$m(z) = \prod \left(z - \bar{\lambda}_j\right).$$

Let $\mathscr{C}_\lambda = \mathscr{F}_\gamma$, where $\gamma = \{1 - yx, m(x), -m(x)\}$. Evidently, $C \in \mathscr{C}_\lambda$ if and only if C is a contraction and $m(C) = 0$.

Define

$$m_j = \frac{\prod_{i \neq j}\left(z - \bar{\lambda}_i\right)}{\prod_{i \neq j}\left(\bar{\lambda}_j - \bar{\lambda}_i\right)},$$

for $j = 0, \dots, n$. The following well known lemma expresses the finite dimensional nature of elements of $\mathscr{C}_\lambda$.

Lemma 4.3. *If X is an operator on a Hilbert space H and $m(X) = 0$, then $m_j(X)$ satisfy*

(i) $m_j(X)^2 = m_j(X)$;
(ii) $m_j(X)m_i(X) = 0$ *if* $i \neq j$;
(iii) $\Sigma m_j(X) = I$; *and*
(iv) $Xm_j(X) = \bar{\lambda}_j m_j(X)$.

Proof: Let $r(z) \doteq \Sigma m_j(z) - 1$. Since r is polynomial of degree n and $r(\bar{\lambda}_j) = 0$ for each $j = 0, 1, \dots, n$, it follows that $r = 0$. Thus,

$$I = \sum m_j(X).$$

Noting that m divides $m_j m_i$ if $i \neq j$ establishes (ii). Combining (ii) and (iii) proves (i). To verify (iv), notice that since $m(X) = 0$, $(X - \bar{\lambda}_j)m_j(X) = 0$. ■

Hereditary polynomials and families are ideal for studying lifting theorems. Suppose $A \in \mathscr{L}(H)$ and $J \in \mathscr{L}(K)$ where K contains H. If H is invariant for J, and $A = J|_H$ we say A *lifts* to J. Geometrically, the condition that A lifts to J can be described as a block operator matrix

$$J = \begin{bmatrix} A & X \\ 0 & Y \end{bmatrix} \tag{4.1}$$

with respect to the orthogonal decomposition of K as $H \oplus (K \ominus H)$. Often H is not explicitly a subspace of K; rather, there exists an isometry $V\colon H \mapsto K$ and we identify H with VH. In this case, the condition that H is an invariant subspace for J and $A = J|_H$ becomes $VA = JV$. For instance, Theorem 3.3 says that a 3-selfadjoint operator lifts to a Jordan operator.

A distinguished member of the class $\mathscr{C}_\lambda$ is (up to unitary equivalence) the $(n+1) \times (n+1)$ matrix E acting on the Hilbert space K with basis the eigenvectors k_j, $j = 0, 1, \dots, n$ satisfying $Ek_j = \overline{\lambda}_j k_j$, and

$$\langle k_j, k_i \rangle = \frac{1}{1 - \overline{\lambda}_j \lambda_i}.$$

Operator theorists know E as the restriction of the adjoint of the unilateral shift to the span of the kernel functions $\{k(\cdot, \lambda_j)\}$, where $k(z, \lambda) = (1 - z\overline{\lambda})^{-1}$ is the Szego kernel.

Let E^α denote the orthogonal direct sum of E with itself α times. The following is a version of the Sz.-Nagy Dilation Theorem **[21, 22]**. The proof indicated below is a version of the deBranges-Rovnyak construction **[9]**.

Theorem 4.4. *If $C \in \mathscr{C}_\lambda \cap \mathscr{L}(H)$, then C lifts to E^α, where α is (at most) the dimension H. If T is in $\mathscr{C} \cap \mathscr{L}(H)$, then T lifts to W^*, where W is an isometry.*

Proof: The second part of the theorem is well known and there are numerous proofs. We refer the reader to **[22]**.

It is possible to obtain the first part of the theorem from the second part. However, we choose to proceed directly giving a construction with the property that if $C \in \mathscr{C}_\lambda \cap L(H)$ and the dimension of H is finite, then all the Hilbert spaces involved in the construction have finite dimension. It will be convenient to represent E^α as a tensor product. Given Hilbert spaces H and H', a Hilbert space $H \otimes H'$ is constructed by defining the inner product of elementary tensors,

$$\langle h \otimes h', k \otimes k' \rangle = \langle h, k \rangle \langle h', k' \rangle,$$

and taking the closure, in the induced norm, of all finite linear combinations of elementary tensors. If $X \in \mathscr{L}(H)$ and $X' \in \mathscr{L}(H')$, we obtain an operator $X \otimes X'$ by defining

$$(X \otimes X')(h \otimes h') = Xh \otimes X'h'.$$

In particular, letting I denote the identity on H, we have $E^\alpha = I \otimes E$; i.e., on elementary tensors,

$$E^\alpha h \otimes k = h \otimes Ek.$$

In particular

$$E^\alpha h \otimes k_j = \overline{\lambda}_j h \otimes k_j.$$

Now let D denote the positive semidefinite square root of $I - C^*C$. Define $V\colon H \mapsto H \otimes K$ by

$$Vh = \sum_j Dm_j(C)h \otimes k_j.$$

A direct computation shows that $\langle Vh, Vh' \rangle = \langle h, h' \rangle$ for all $h, h' \in H$. Thus V is an isometry. Moreover, $VC = (I \otimes E)V$. ■

Given a family $\mathscr{F}$ and some $A \in \mathscr{F}$, we would like to find a $J \in \mathscr{F}$, distinguished in some way, such that A lifts to J. It is also reasonable to expect that there are certain elements of $\mathscr{F}$ that we cannot lift in a nontrivial way. An operator $A \in \mathscr{F}$ is called *extremal* if whenever A lifts to $J \in \mathscr{F}$ as in (4.1), it follows that $X = 0$. Algebraically, A is extremal if whenever there exists a $J \in \mathscr{F}$ and isometry V such that $VA = JV$, it follows that $VA^* = J^* V$.

Theorem 4.5. *$A \in \mathscr{C}_\lambda$ is extremal in $\mathscr{C}_\lambda$ if and only if $A = E^\alpha$ for some α. $T \in \mathscr{C}$ is extremal if and only if T^* is an isometry.*

Proof: We give a proof of the second part only, since it is somewhat easier than the first part, it illustrates the basic approach, and, while it is known, is not usually expressed in this way.

Suppose T^* is an isometry and

$$J = \begin{bmatrix} T & X \\ 0 & Y \end{bmatrix}$$

is in $\mathscr{C}$. Since J^* is then in $\mathscr{C}$, it follows that

$$0 \leq I - JJ^* = \begin{bmatrix} I - TT^* - XX^* & -XY^* \\ -YX^* & I - YY^* \end{bmatrix}.$$

Since $I - TT^* = 0$, it follows that $X = 0$. We conclude that T is extremal.

Conversely, suppose T^* is not an isometry. Then there exists a nonzero vector x such that $I - TT^* - xx^* \geq 0$. Let

$$J = \begin{bmatrix} T & x \\ 0 & 0 \end{bmatrix}.$$

Compute

$$I - JJ^* = \begin{bmatrix} I - TT^* - xx^* & 0 \\ 0 & 0 \end{bmatrix},$$

from which it follows that $J \in \mathscr{C}$. Hence in this case T is not extremal. ∎

Agler showed in [2] that if $\mathscr{F}$ is a family and $A \in \mathscr{F}$, then there exists $J \in \mathscr{F}$ extremal such that A lifts to J. A consequence is the following theorem, which is a cornerstone of Agler's abstract approach to model theory. A subspace M of the Hilbert space H is *reducing* for $J \in \mathscr{L}(H)$ if M is invariant for both J and J^*. In this case J decomposes as a direct sum $J_1 \oplus J_2$ with respect to the orthogonal sum $H = M \oplus (H \ominus M)$.

Theorem 4.6. *Given a family $\mathscr{F}$, there exists a subcollection $\partial\mathscr{F}$ of $\mathscr{F}$ such that*

(i) *$\partial\mathscr{F}$ is closed with respect to orthogonal direct sums;*
(ii) *$\partial\mathscr{F}$ is closed with respect to unital $*$-representations;*
(iii) *$\partial\mathscr{F}$ is closed with respect to restriction to reducing subspaces; i.e., if $J \in \mathscr{L}(H) \cap \partial\mathscr{F}$ and M is a reducing subspace for J, then $J|_M \in \partial\mathscr{F}$;*
(iv) *For each $A \in \mathscr{F}$, there exists $J \in \partial\mathscr{F}$ such that A lifts to J.*
(v) *If $\mathscr{B}$ is any subcollection of $\mathscr{F}$ satisfying (i), (ii), (iii), and (iv) above, then $\partial\mathscr{F} \subset \mathscr{B}$.*

Any collection $\mathscr{B}$ satisfying the conditions (i), (ii), (iii), and (iv) of Theorem 4.6 is called a *model* for $\mathscr{F}$. The collection $\partial\mathscr{F}$ is the *boundary* (often called the *Agler boundary*) of the family $\mathscr{F}$. The boundary of $\mathscr{F}$ always contains the extremal

elements of $\mathscr{F}$ and in many cases the extremals in $\mathscr{F}$ constitute the boundary of $\mathscr{F}$. For example, Theorem 4.4 says that $\mathscr{E}_\lambda = \{E^\alpha\colon \alpha\}$ are the extremals of $\mathscr{E}_\lambda$. On the other hand, $\mathscr{E}_\lambda$ obviously satisfies (i), Theorem 4.4 implies that it satisfies (iv), and the proof of Theorem 4.4 shows that it satisfies (ii) and (iii). Thus $\mathscr{E}_\lambda = \partial\mathscr{E}_\lambda$. Similar reasoning shows that the boundary of $\mathscr{C}$ is the collection of adjoints of isometries, commonly called *coisometries*.

Description of the boundary and of the extremal elements for a given family is a fundamental problem in the theory and applications of hereditary classes. So far, this problem has been satisfactorily solved only for very few (but important) families.

The general theory of hereditary classes, families, their boundaries and extremal elements, provides a paradigm for several important special classes in operator theory: (a) isometries and unitaries (Wold decomposition); (b) contraction and co-isometries (Theorems 4.4 and 4.5); (c) contractive subnormal and contractive normal operators (see **[11]** for a thorough account of the theory and applications of subnormal operators); (d) numerical radius contractions (suggested by the recent work **[13]**).

There are many open problems in the theory of hereditary classes, both concrete and of general nature. We would like to emphasize here problems concerning special features that matrices (i.e., operators on finite dimensional Hilbert spaces) might have. For example:

Problem 4.7. *Is there a simple condition (say, on the hereditary polynomials that determine the family) that guarantees that the matrices in the boundary are automatically extremal? More generally, when does the set of extremals coincide with the boundary?*

Problem 4.8. *Suppose $\mathscr{F}$ is a family with the boundary $\partial\mathscr{F}$. When does every matrix in $\mathscr{F}$ lift to a matrix in $\partial\mathscr{F}$?*

We end this section by returning to the 3-selfadjoint operators. Recall that the class of 3-selfadjoint operators is not a family, since it is not bounded. However, by imposing some canonical normalization we do obtain a family. Let $p(x, y)$ denote the hereditary polynomial $-\frac{1}{2}(y-x)^2$ and fix an interval $[a, b]$. Then the set of 3-selfadjoint operators A with spectrum in a given interval $[a, b]$ and such that

$$p(A) \leq c^2$$

is a family that we denote by $\mathscr{S} = \mathscr{S}_{[a,b],c}$. If $J = T + N$ is a Jordan operator, then direct computation shows that $J \in \mathscr{S}$ if and only if the spectrum of J is in $[a, b]$ and $N^*N \leq c^2$.

Theorem 4.9. *The boundary $\partial\mathscr{S}$ of $\mathscr{S}$ consists of precisely those Jordan operators $J = T + N \in \mathscr{S}$, where*

$$N^*N + NN^* = c^2 I.$$

Moreover, each $J \in \partial\mathscr{S}$ is extremal.

If $J \in \mathscr{S} = \mathscr{S}_{[a,b],c}$ is a matrix, then it is possible to give a simple construction of the lift of J to a matrix J_0 in the boundary of $\mathscr{S}$. For simplicity, assume that $c = 1$ and that J is completely nonselfadjoint, i.e., there is no nontrivial subspace that is invariant for both J and J^* and on which J is selfadjoint. Then J is (up to

unitary similarity) an orthogonal sum of matrices of the form

$$J = \begin{bmatrix} tI & X \\ 0 & tI \end{bmatrix} \tag{4.2}$$

where $\|X\| \le c = 1$, and $t \in [a, b]$ (cf. Theorem 2.7). The complete nonselfadjointness of J guarantees that X is onto. Let $D = \sqrt{I - X^*X}$, the *defect* of X. Define

$$V\colon \mathbf{C}^p \oplus \mathbf{C}^p \to \mathbf{C}^p \oplus \mathbf{C}^p \oplus \mathbf{C}^p \oplus \mathbf{C}^p,$$

where $p \times p$ is the size of I in (4.2), by

$$V \begin{bmatrix} g \\ f \end{bmatrix} = \begin{bmatrix} Df \\ 0 \\ g \\ Xf \end{bmatrix}.$$

Let

$$J_0 = \begin{bmatrix} tI & I & 0 & 0 \\ 0 & tI & 0 & 0 \\ 0 & 0 & tI & I \\ 0 & 0 & 0 & tI \end{bmatrix}.$$

One verifies that V is an isometry, and $VJ = J_0 V$. Clearly, J_0 is in the boundary of $\mathscr{S}$ as described in Theorem 4.9.

As a final remark we note that an operator T satisfying

$$T^{*n}T^n = I + \frac{n}{2}(-3 + 4T^*T - T^{*2}T^2) + \frac{n^2}{2}(I - 2T^*T + T^{*2}T^2)$$

is called a 3-isometry. If T is 3-selfadjoint, then $\exp(isT)$ is a 3-isometry for real s. The notion of 3-isometry can be generalized in a natural way to m-isometry, by requiring that $T^{*n}T^n$ is a polynomial of degree $m - 1$ as a function of n. It turns out that there exist 2-isometries that are not isometries, in contrast to the situation for 2-selfadjoint operators. Very important advances in the theory of classes of m-isometries, including applications to differential operators, disconjugacy, and Brownian motion have been obtained recently in [**4**].

5. SELFADJOINT OPERATORS AND RELATED CLASSES IN KREIN SPACES. Let H be a Hilbert space (over the complex numbers), and let J be a selfadjoint operator on H such that $J^2 = I$. Consider the sesquilinear form $[\cdot, \cdot]$ induced by J:

$$[x, y] = \langle Jx, y\rangle, \qquad x, y \in H,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in H. The corresponding quadratic form $[x, x]$ is indefinite (unless $J = I$ or $J = -I$), in other words, there exist $x, y \in H$ for which $[x, x] < 0$ and $[y, y] > 0$. The space H, together with the sesquilinear form $[\cdot, \cdot]$ generated by some J, as just described, is called a *Krein space*. One can also define the Krein spaces intrinsically, by starting with a topological vector space and a continuous sesquilinear form on it, and by imposing suitable completeness and nondegeneracy axioms.

The theory of operators in Krein spaces has been extensively developed in the recent half century, largely motivated by application in the physical sciences (description of physical processes governed by boundary value problems, partial differential equations, etc.). Recently, there is a renewed interest in Krein space

theory, mainly due to numerous new applications, especially in modern electrical engineering; for example, applications of finite dimensional Krein spaces to basic optimal worst case control problems in linear systems are described in [7]. Several books are dedicated to the theory of operators in Krein spaces and its applications [**5**, **6**, **10**, **14**, **18**]; the book [**14**] deals exclusively with finite-dimensional Krein spaces, and as such is more accessible as an introduction than the other four books.

Here we concern ourselves with hereditary classes in Krein spaces. As in the Hilbert space case, such classes are defined using polynomial operator identities of type (1.3), where, of course, A^* stands now for the Krein space adjoint:

$$[Ax, y] = [x, A^* y] \qquad \text{for all } x, y \in H.$$

(Everywhere in this section the adjoint is understood in the Krein space sense.) The works in this area we are aware of deal mainly with dilation of isometries (the paper [**12**] and the extensive bibliography there are a good source concerning this theory); a possibility of developing some results on 3-selfadjoint operators in Krein spaces and their potential connections to the theory of Sturm-Liouville operators was indicated already in [**15**]. There does not exist even a partial theory of hereditary classes and families of Krein space operators, analogous in spirit to the basic results presented in Section 3. Development of such a theory appears to us a worthwhile research project. One should be warned, though: Many facts taken for granted in the Hilbert space situation are hopelessly lost in Krein spaces. For example, a selfadjoint operator in Krein space need not have real spectrum; a selfadjoint operator in finite-dimensional Krein space need not be diagonalizable.

This section can also be regarded as a rather unusual introduction to some aspects of Krein spaces, which is motivated by the notion of 2-selfadjoint operators and related hereditary classes. Namely, we consider the class of operators A that satisfy

$$A^2 - 2A^*A + A^{*2} = 0, \tag{5.1}$$

(naturally, such A will be termed *2-selfadjoint operators*), and the classes of operators A that satisfy the related identities

$$A^*A = A^2; \tag{5.2}$$

$$AA^* = A^2; \tag{5.3}$$

$$A^* = A \quad \text{(selfadjoint operators)}. \tag{5.4}$$

It turns out that the four classes of operators (defined by (5.1), (5.2), (5.3), and (5.4), respectively) are all distinct in the Krein space framework, in contrast with the Hilbert spaces (see Example 5.2 below).

We assume from now on that the Krein space defined by J, as in the beginning of this section, is such that either $J = I$, or $-1 \in \sigma(J)$ and -1 is an eigenvalue of J having finite multiplicity. Such Krein spaces are commonly called *Pontryagin spaces*. The total multiplicity of -1 as an eigenvalue of J is a finite number, called the *defect* of the Pontryagin space. When $J = I$, we formally define the defect to be zero.

For an operator A on H, denote by A_R (respectively, A_J) the selfadjoint (respectively, skew-adjoint) part of A. These operators are uniquely defined by the properties that A_R and A_J are selfadjoint in the Pontryagin space sense, i.e.,

$$[A_R x, y] = [x, A_R y], [A_J x, y] = [x, A_J y]$$

for all $x, y \in H$, and $A = A_R + iA_J$. The following result was proved in [**20**].

Theorem 5.1. *Assume H is a Pontryagin space with defect q, and let A be an operator on H such that*

$$A^2 - 2A^*A + A^{*2} = 0. \tag{5.5}$$

Then $\sigma(A_J) = \{0\}$, rank $A_J \leq 2q$, and

$$\operatorname{rank}(A^*A - AA^*) \leq 2q - 1 \qquad \text{if } q > 0. \tag{5.6}$$

Theorem 3.1 is a particular case of Theorem 5.1 (where one takes $q = 0$).

We illustrate Theorem 5.1, as well as some other aspects of the theory of Krein space operators, by the following example.

Example 5.2. Let $H = \mathbf{C}^3$ (with the standard Hilbert space structure), and let $J = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$. Note that H has defect one as a Pontryagin space. Consider

$$A = \begin{bmatrix} 0 & p & r \\ 0 & 0 & q \\ 0 & 0 & 0 \end{bmatrix},$$

where p, q, and r are complex numbers. Using the formula

$$A^* = J(\overline{A})^T J,$$

where $(\overline{A})^T$ stands for the conjugate transpose of A, we obtain

$$A^* = \begin{bmatrix} 0 & \bar{q} & \bar{r} \\ 0 & 0 & \bar{p} \\ 0 & 0 & 0 \end{bmatrix}.$$

Thus, $A = A^*$ if and only if $p = \bar{q}$ and r is real. We see immediately that there exist selfadjoint nondiagonalizable matrices. Computations show that (5.2) holds if and only if $q(p - \bar{q}) = 0$, (5.3) holds if and only if $p(p - \bar{q}) = 0$, and (5.1) holds if and only if pq is real. Thus, in this example, all four classes defined by (5.1), (5.2), (5.3), and (5.4) are distinct. ■

Several interesting open problems present themselves naturally in the context of the material of this section. For example:

Problem 5.3. *Describe the set of operators in a finite dimensional Krein space that satisfy*

$$A^2 - 2A^*A + A^{*2} = 0, \qquad AA^* = A^2.$$

Problem 5.4. *Determine the indecomposables (analogous to those in the Hilbert space situation, see Section 2) in the class of 2-selfadjoint operators in finite dimensional Krein spaces.*

ACKNOWLEDGMENTS. This work is partially supported by the NSF Grants DMS-9307966 (for SAM) and DMS-9500924 (for LR). An elementary proof of the fact that (1.1) implies that the matrix A is Hermitian was communicated to us by L. DeAlba.

REFERENCES

1. J. Agler, Sub-Jordan operators, Bishop's theorem, Spectral inclusion and spectral sets, *J. Operator Theory* 7 (1982), 373–395.
2. J. Agler, An Abstract Approach to Model Theory, in: J. B. Conway, B. Morrel, *Survey of Some Recent Results in Operator Theory*, Vol. **II**, Pitman/92 (1988), 1–24.

3. J. Agler, A disconjugacy theorem for Toeplitz operators, *American J. of Math.* **112** (1990), 1–14.
4. J. Agler and M. Stankus, m-isometric transformations of Hilbert space, I, *Integral Equations and Operator Theory* **21** (1995), 383–429; II, *Integral Equations and Operator Theory* **23** (1995), 1–48; III, preprint.
5. T. Ando, *Linear Operators on Krein Spaces*, Lecture Notes, Sapporo, Japan, 1979.
6. T. Ya. Azizov and I. Iokhvidov, *Linear Operators in Spaces with an Indefinite Metric*, J. Wiley and Sons, New York, 1989 (Translated from Russian).
7. J. A. Ball, I. Gohberg and L. Rodman, *Interpolation of Rational Matrix Functions*, Operator Theory: Advances and Applications, Vol. 45, Birkhäuser, Basel, 1990.
8. J. A. Ball and J. W. Helton, Nonnormal dilations, disconjugacy and constrained spectral factorization, *Integral Equations and Operator Theory* **3** (1980), 216–309.
9. L. de Branges and J. Rovnyak, Canonical models in quantum scattering theory, pp. 295–392 in: *Perturbation Theory and its Applications in Quantum Mechanics* (C. H. Wilcox, ed.), Wiley, New York, London, Sidney, 1966.
10. J. Bognar, *Indefinite Inner Product Spaces*, Springer-Verlag, Berlin, 1974.
11. J. B. Conway, *The Theory of Subnormal Operators*, Math. Surveys and Monographs, Vol. 36, American Mathematical Society, Providence, R. I., 1991.
12. M. A. Dritschel and J. Rovnyak, Extension theorems for contraction operators in Krein spaces, *Operator Theory: Advances and Applications* (I. Gohberg, ed.) Vol. 47, pp. 221–305, Birkhäuser Verlag, 1990.
13. M. A. Dritschel and H. J. Woerdeman, The Agler boundary and linear extreme points in the numerical radius norm, *Memoirs of Amer. Math. Soc.*, to appear.
14. I. Gohberg, P. Lancaster, and L. Rodman, *Matrices and Indefinite Scalar Products*, Operator Theory: Advances and Applications, Vol. 8, Birkhäuser, Basel, 1983.
15. J. W. Helton, Infinite dimensional Jordan operators and Sturm-Liouville conjugate point theory, *Trans. Amer. Math. Soc.*, **170** (1972), 305–331.
16. J. W. Helton, Jordan operators in infinite dimensions and Sturm-Liouville conjugate point theory, *Bull. Amer. Math. Soc.*, **78** (1972), 57–61.
17. R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.
18. I. S. Iohvidov, M. G. Krein, and H. Langer, *Introduction to the Spectral Theory of Operators in Spaces with an Indefinite Metric*, Akademie Verlag, Berlin, 1982.
19. K. R. Laberteux, Problem 10377, *Amer. Math. Monthly* **101** (1994), 362.
20. S. A. McCullough and L. Rodman, Two-selfadjoint operators in Krein spaces, *Integral Equations and Operator Theory*, **26** (1996), 202–209.
21. B. Sz.-Nagy and C. Foiaş, Sur les contractions de l'espace de Hilbert, *Acta Sci. Math.* **25** (1964), pp. 38–71.
22. B. Sz.-Nagy and C. Foiaş, *Harmonic Analysis of Operators on Hilbert Space*, American Elsevier Publishing, New York, 1970.

Scott A. McCullough
Department of Mathematics
University of Florida
Gainesville, FL 32611
sam@math.ufl.edu

Leiba Rodman
Department of Mathematics
The College of William and Mary
Williamsburg, VA 23187-8795
lxrodm@math.wm.edu

We have college students come in and study certain aspects of the show. We even had a mathematician come in and figure out the best strategy for spinning the wheel.

Bob Barker, quoted in "Who's Saying 'The Price is Right'?"
The Salt Lake Tribune, February 9, 1997, p. E1

Major Centers of Triangles

Clark Kimberling

Not long ago, the *Monthly* carried an article with a most intriguing title: "The Rise, Fall, and Possible Transfiguration of Triangle Geometry: A Mini-history" [**4**]. Among the possible transfigurations mentioned was a recognition of triangle centers as functions instead of mere points. Here, we wish to pose certain problems, with partial solutions, for which one is forced into the functional mode —otherwise the problems do not make sense.

To get started, consider the following: Let $x(ABC)$ denote a procedure for constructing a center X of $\triangle ABC$, such as its incenter I, and let $y(ABC)$ denote a procedure for finding another (perhaps the same) center Y of $\triangle ABC$, such as its centroid G. For a given choice of X we can form the three triangles XBC, XCA, XAB, and apply the function y to each of these to find their centers:

$$y(\triangle XBC) = Y_A, \qquad y(\triangle AXC) = Y_B, \qquad y(\triangle ABX) = Y_C.$$

Let

$$A' = XY_A \cap BC, \qquad B' = XY_B \cap CA, \qquad C' = XY_C \cap AB.$$

> *Question A:* For a given function x, what functions y yield a triangle $A'B'C'$ that is perspective to $\triangle ABC$, i.e., is such that the lines AA', BB', CC' concur in a point Z?

It is easy to check that if x constructs the incenter and y constructs the centroid, then the resulting $\triangle A'B'C'$ is perspective to $\triangle ABC$, and Z is the centroid of $\triangle ABC$. Similarly, if y constructs the circumcenter, then Z is just the original center, I, as shown in Figure 1.

One can turn these geometric procedures into analytic ones via trilinear coordinates and a general definition of "center." We shall show that if x constructs the incenter, the analysis results in an unexpectedly simple form for a particular class of centers called major centers.

1. TRILINEARS. The position of any point P in the plane of $\triangle ABC$ is uniquely determined by any three numbers α, β, γ that are proportional to the directed distances from P to sidelines BC, CA, AB. For example, the incenter, I, being equidistant from the sidelines, is given by $\alpha = \beta = \gamma = 1$, and we write $I = 1:1:1$. Similarly,

$$\begin{aligned}
&\text{centroid} = \csc A : \csc B : \csc C,\\
&\text{circumcenter} = \cos A : \cos B : \cos C,\\
&\text{orthocenter} = \sec A : \sec B : \sec C,\\
&\text{nine-point center} = \cos(B - C) : \cos(C - A) : \cos(A - B).
\end{aligned}$$

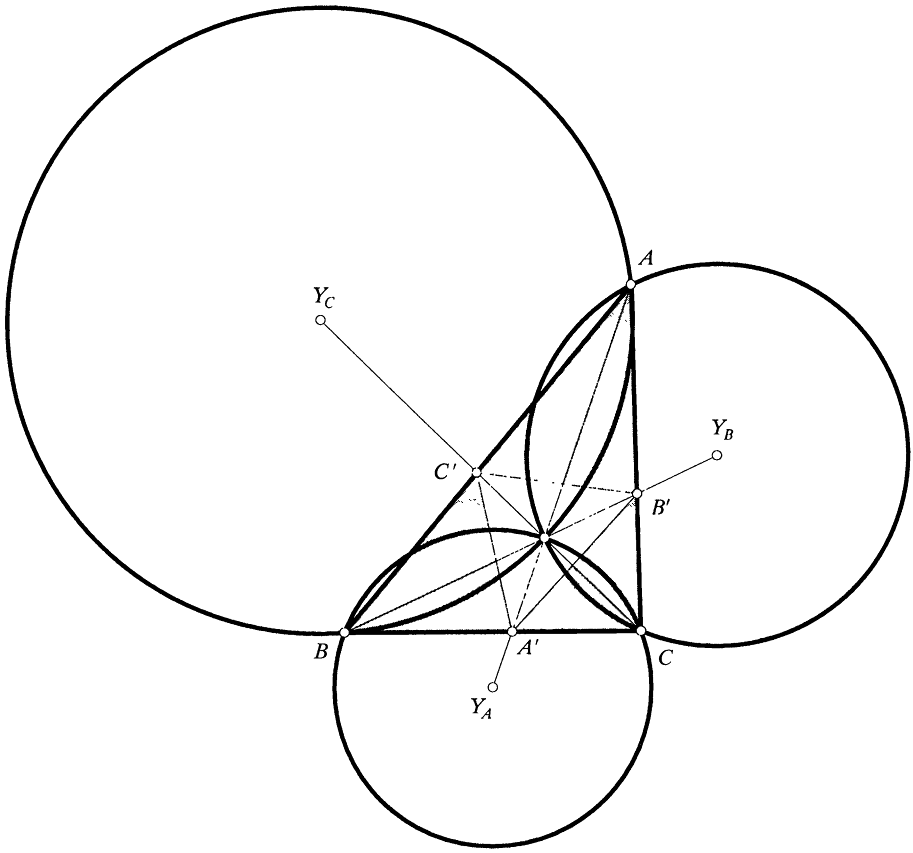


Figure 1. ($X = I$ and $Y =$ circumcenter) $\Rightarrow$ ($\triangle\ A'B'C'$ perspective to $\triangle\ ABC$).

Three points $P_i = \alpha_i\colon \beta_i\colon \gamma_i$ are collinear if and only if

$$\begin{vmatrix} \alpha_1 & \beta_1 & \gamma_1 \\ \alpha_2 & \beta_2 & \gamma_2 \\ \alpha_3 & \beta_3 & \gamma_3 \end{vmatrix} = 0. \tag{1}$$

Replacing $\alpha_1\colon \beta_1\colon \gamma_1$ by a variable point $\alpha\colon \beta\colon \gamma$ in (1) gives an equation for the line P_2P_3. The dual to the collinearity expressed by (1) is the concurrence of three lines $\alpha\alpha_i + \beta\beta_i + \gamma\gamma_i = 0$, $i = 1, 2, 3$. Extending the duality, the intersection of the last two of these lines is the point

$$\beta_2\gamma_3 - \gamma_2\beta_3\colon \gamma_2\alpha_3 - \alpha_2\gamma_3\colon \alpha_2\beta_3 - \beta_2\alpha_3. \tag{2}$$

For more about trilinears, see the recent treatment by Coxeter [**2**] or Oldknow [**9**]. Boyer [**1**] traces the origin of trilinear coordinates back to Möbius.

When trilinears for points are written in terms of the vertex angles A, B, C, the points may be regarded as functions of the variables A, B, C. We take this point of view and note three particular properties of traditional special points of a triangle: suppose a point P has trilinears

$$f(A,B,C)\colon g(A,B,C)\colon h(A,B,C)$$

satisfying

(i) $g(A,B,C) = f(B,C,A)$ and $h(A,B,C) = f(C,A,B)$;

(ii) $f(A,C,B) = f(A,B,C)$;

(iii) if P is written as $u(a,b,c)\colon u(b,c,a)\colon u(c,a,b)$, where a,b,c are the sidelengths of $\triangle ABC$, then u is homogeneous in the variables a,b,c. (By the law of sines and property (i), such a u must exist.)

Then P is called (in **[6]** and **[8]**) a *triangle center*, or simply a *center*. It is easy to verify that the five special points already mentioned are centers. This is true also for points named after Lemoine, Gergonne, Nagel, Spieker, Feuerbach, Fermat, Napoleon, Euler, Steiner, Hofstadter, Parry, Yff, and others. We now define $\mathcal{F}$ to be the set of centers and recast Question A as follows:

Problem A: Given $X =$ incenter, find all Y in $\mathcal{F}$ for which $\triangle A'B'C'$ is perspective to $\triangle ABC$.

In order to evaluate Y in the three triangles, we shall form trilinears for Y_A relative to $\triangle XBC$, for Y_B relative to $\triangle AXC$, and for Y_C relative to $\triangle ABX$, and then transform these to trilinears relative to the reference triangle ABC.

2. COORDINATE TRANSFORMATIONS. Any three noncollinear points

$$P_i = f_i(a,b,c) : g_i(a,b,c) : h_i(a,b,c) : \quad i = 1,2,3,$$

determine a triangle with vertices P_1, P_2, P_3. The triangle can be represented as a matrix

$$M = \begin{pmatrix} f_1(a,b,c) & g_1(a,b,c) & h_1(a,b,c) \\ f_2(a,b,c) & g_2(a,b,c) & h_2(a,b,c) \\ f_3(a,b,c) & g_3(a,b,c) & h_3(a,b,c) \end{pmatrix}.$$

Let F_i, G_i, H_i denote the functions satisfying

$$M^{-1} = \frac{1}{|M|}\begin{pmatrix} F_1(a,b,c) & G_1(a,b,c) & H_1(a,b,c) \\ F_2(a,b,c) & G_2(a,b,c) & H_2(a,b,c) \\ F_3(a,b,c) & G_3(a,b,c) & H_3(a,b,c) \end{pmatrix}.$$

Let $\alpha' : \beta' : \gamma'$ be trilinears *relative to M* for a point Y. Then the matrix equation

$$\begin{pmatrix} \alpha & \beta & \gamma \end{pmatrix} = \begin{pmatrix} \alpha' & \beta' & \gamma' \end{pmatrix} DM \tag{3}$$

gives trilinears $\alpha : \beta : \gamma$ *relative to* $\triangle ABC$ for the point Y, where

$$D = \begin{pmatrix} \delta_1 D_1 & 0 & 0 \\ 0 & \delta_2 D_2 & 0 \\ 0 & 0 & \delta_3 D_3 \end{pmatrix},$$

$$D_1 = \sqrt{F_1^2 + F_2^2 + F_3^2 - 2F_2F_3\cos A - 2F_3F_1\cos B - 2F_1F_2\cos C}\,,$$

$$D_2 = \sqrt{G_1^2 + G_2^2 + G_3^2 - 2G_2G_3\cos A - 2G_3G_1\cos B - 2G_1G_2\cos C}\,,$$

$$D_3 = \sqrt{H_1^2 + H_2^2 + H_3^2 - 2H_2H_3\cos A - 2H_3H_1\cos B - 2H_1H_2\cos C}\,,$$

and $\delta_i = 1$ for $i = 1, 2, 3$. For a proof, see [**6**]. In order to say when δ_i is -1 and when $+1$, we must first define the *positive side* of a sideline of a triangle as the set of points in the plane of the triangle that lie on the same side of the sideline as does the vertex that is not already on the sideline. The *negative side* is the other side, and the *directed distance* from a point to a sideline is positive or negative according to which side the point lies on. Now, the directed distance from Y to sideline P_2P_3 is $\delta_1 h_1/D_1$, where $h_1 = \alpha F_1 + \beta F_2 + \gamma F_3$, and δ_1 is determined as follows: if Y lies on the positive side of P_2P_3 and $h_1 < 0$, then $\delta_1 = -1$; if Y lies on the negative side of P_2P_3 and $h_1 > 0$, then $\delta_1 = -1$; otherwise, $\delta_1 = 1$. The numbers δ_2 and δ_3 are defined analogously.

Returning to Problem A, we eventually want to allow X to be points other than the incenter, so we write $X = x\colon y\colon z$. The relevant matrices are

$$M = \begin{pmatrix} x & y & z \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } M^{-1} = \frac{1}{x^3}\begin{pmatrix} 1 & -y & -z \\ 0 & x & 0 \\ 0 & 0 & x \end{pmatrix},$$

so that the transformation (3) can be written out as

$$\alpha\colon \beta\colon \gamma = x\delta_1 D_1 \alpha'\colon y\delta_1 D_1 \alpha' + \delta_2 D_2 \beta'\colon z\delta_1 D_1 \alpha' + \delta_3 D_3 \gamma'. \tag{4}$$

This equation shows how to write trilinears $\alpha\colon \beta\colon \gamma$ for Y_A relative to $\triangle\, ABC$ in terms of trilinears $\alpha'\colon \beta'\colon \gamma'$ for Y_A given relative to $\triangle\, XBC$. Of course, $\alpha'\colon \beta'\colon \gamma'$ must be given in terms of the sidelengths of $\triangle\, XBC$, which in turn depend on the choice of point X. Perhaps the simplest case is $X =$ incenter.

3. MAJOR CENTERS AS SOLUTIONS OF PROBLEM A.

For Problem A, we have $X = 1:1:1$ and $\delta_1 = \delta_2 = \delta_3 = 1$ in equation (4), and, with sidelengths of $\triangle\, XBC$ as labeled in Figure 2, we obtain from equation (4) the transformation

$$\alpha\colon \beta\colon \gamma = \alpha'\colon \alpha' + 2\beta' \cos(C/2)\colon \alpha' + 2\gamma' \cos(B/2).$$

It is now easy to see that lines XY_A and BC meet in the point

$$A' = 0\colon \beta' \sec(B/2)\colon \gamma' \sec(C/2).$$

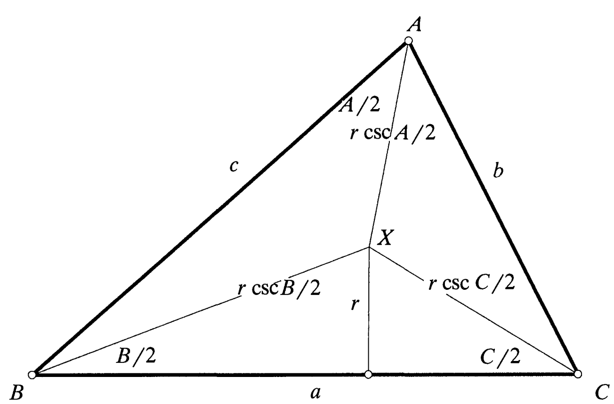


Figure 2. Incenter, X, inradius, r, and associated angles and lengths.

We wish to do the same thing relative to the other two interior triangles, AXC and ABX, obtaining points B' and C', then to find equations for the three lines AA', BB', CC', and then to find a necessary and sufficient condition for these three lines to concur. Write Y_A as $\alpha_1': \beta_1': \gamma_1'$, and similarly, $Y_B = \alpha_2': \beta_2': \gamma_2'$ and $Y_C = \alpha_3': \beta_3': \gamma_3'$. There are simple equations for the three lines

$$AA': \quad (-1/\beta_1')\cos(B/2)\beta + (1/\gamma_1')\cos(C/2)\gamma = 0, \tag{5}$$

$$BB': \quad (1/\alpha_2')\cos(A/2)\alpha + (-1/\gamma_2')\cos(C/2)\gamma = 0, \text{ and} \tag{6}$$

$$CC': \quad (1/\alpha_3')\cos(A/2)\alpha + (-1/\beta_2')\cos(B/2)\beta = 0. \tag{7}$$

Using equation (1), we conclude that these lines concur if and only if

$$\alpha_3' \beta_1' \gamma_2' = \alpha_2' \beta_3' \gamma_1', \tag{8}$$

where

$$\begin{aligned}
\alpha_1' &= \alpha'(a, r\csc(C/2), r\csc(B/2)), \; \beta_1' = \alpha'(r\csc(C/2), r\csc(B/2), a),\\
&\qquad \gamma_1' = \alpha'(r\csc(B/2), a, r\csc(C/2)),\\
\alpha_2' &= \alpha'(r\csc(C/2), b, r\csc(A/2)), \; \beta_2' = \alpha'(b, r\csc(A/2), r\csc(C/2)),\\
&\qquad \gamma_2' = \alpha'(r\csc(A/2), r\csc(C/2), b),\\
\alpha_3' &= \alpha'(r\csc(B/2), r\csc(A/2), c), \; \beta_3' = \alpha'(r\csc(A/2), c, r, \csc(B/2)),\\
&\qquad \gamma_3' = \alpha'(c, r\csc(B/2), r\csc(A/2)),
\end{aligned}$$

and r denotes the inradius of $\triangle ABC$. These equations are amenable if Y is a center for which there exist trilinears $\alpha: \beta: \gamma$ such that α is a function of A alone; that is, $\alpha = f(A)$. We call such a center Y a *major center*.

Theorem. *Every major center Y solves Problem A. If $Y = f(A): f(B): f(C)$, then the point of concurrence is the major center given by*

$$f(A/2)\sec(A/2) : f(B/2)\sec(B/2) : f(A/2)\sec(C/2). \tag{9}$$

Proof: Given $Y = f(A): f(B): f(C)$, we find Y_A, Y_B, Y_C given by

$$\begin{pmatrix} \alpha_1' & \beta_1' & \gamma_1' \\ \alpha_2' & \beta_2' & \gamma_2' \\ \alpha_3' & \beta_3' & \gamma_3' \end{pmatrix} = \begin{pmatrix} f((\pi - A)/2) & f(B/2) & f(C/2) \\ f(A/2) & f((\pi - B)/2) & f(C/2) \\ f(A/2) & f(B/2) & f((\pi - C)/2) \end{pmatrix}. \tag{10}$$

Then equation (8) holds, and equations (5) to (7) together with (2) imply

$$BB' \cap CC' = -\frac{\cos(B/2)\cos(C/2)}{\gamma_2' \beta_3'} : -\frac{\cos(C/2)\cos(A/2)}{\gamma_2' \alpha_3'} : -\frac{\cos(A/2)\cos(B/2)}{\alpha_2' \beta_3'},$$

so that (10) leads directly to (9). ■

For example, if $Y = \sin A: \sin B: \sin C$ (the *symmedian point*, or *Lemoine point*), then

$$\begin{pmatrix} \alpha_1' & \beta_1' & \gamma_1' \\ \alpha_2' & \beta_2' & \gamma_2' \\ \alpha_3' & \beta_3' & \gamma_3' \end{pmatrix} = \begin{pmatrix} \cos(A/2) & \sin(B/2) & \sin(C/2) \\ \sin(A/2) & \cos(B/2) & \sin(C/2) \\ \sin(A/2) & \sin(B/2) & \cos(C/2) \end{pmatrix},$$

and $BB' \cap CC' = -(\cos(B/2)\cos(C/2))/(\sin(C/2)\cos(B/2))$, so that the point of concurrence (9) is $\tan(A/2): \tan(B/2): \tan(C/2)$.

It can be confirmed by *Mathematica* using trilinears that certain centers other than major centers also solve Problem A. Among them are the center of the *Kiepert hyperbola*, given by $\alpha = \sin A \sin^2(B - C)$, and the center of the *Jerabek hyperbola*, given by $\alpha = \cos A \sin^2(B - C)$. The Kiepert hyperbola was recently resurrected in **[4]**, and the Jerabek, in **[10]**. One wonders if there is a nice way to write the *general* solution of Problem A.

4. MAJOR CENTERS OBTAINED AS LIMITS. Let $F(A)$ denote the infinite product

$$\sec(A/2)\sec(A/4)\sec(A/8)\sec(A/16)\sec(A/32)\cdots,$$

and let $Z = F(A)\colon F(B)\colon F(C)$. It is a charming little calculus problem to prove that the product converges for all real numbers A to the limit $F(A) = \sin(2A)/2A$, so that Z is the center

$$\sin(2A)/A : \sin(2B)/B : \sin(2C)/C,$$

of interest because the vertex angles A, B, C appear in the denominators without being the arguments of trigonometric functions. These angles also appear "exposed" in another, quite different manner, in which we do not need to know that $F(A) = \sin(2A)/2A$. Starting with any major center $Y = f(A)\colon f(B)\colon f(C)$, iterate (9) to obtain the center

$$\lim_{n\to\infty} \left(f(A/2^n)F(A) : f(B/2^n)F(B) : f(C/2^n)F(C) \right). \tag{11}$$

Suppose now that $Y =$ centroid, so that $f =$ cosecant. Then the limit (11) is the center

$$(1/A)F(A) : (1/B)F(B) : (1/C)F(C). \tag{12}$$

At this point, we recall a kind of conjugate that pervades triangle geometry. Suppose $P = x\colon y\colon z$ is a point not on a sideline of $\triangle ABC$. The reflections of lines AP, BP, CP about the interior bisectors of angles A, B, C respectively, concur in the *isogonal conjugate* of P, given by trilinears $1/x\colon 1/y\colon 1/z$.

Thus, the center in (12) has isogonal conjugate $A/F(A)\colon B/F(B)\colon C/F(C)$. Let $g(A) = A/F(A)$, and using g instead of f, iterate (9) again, obtaining as a limit the center $1/A : 1/B : 1/C$, whose isogonal conjugate is the *bare-angle center* $A\colon B\colon C$. Thus, we have a "construction" of an extremely easy-to-represent "transcendental" center. This mysterious center is mentioned in connection with other transcendental centers conceived a few years ago by Douglas Hofstadter **[7]**.

5. ANOTHER PROBLEM. Returning to Question A, suppose X need not be the incenter but is free to range about the interior of $\triangle ABC$. Further, suppose Y represents an arbitrary center, and as before,

$$Y_A = Y(\triangle XBC),\quad Y_B = Y(\triangle AXC),\quad Y_C = Y(\triangle ABX),$$
$$A' = XY_A \cap BC,\quad B' = XY_B \cap CA,\quad C' = XY_C \cap AB.$$

Problem B: For what choices of Y must AA', BB', CC' concur for *all* choices of X?

It is easy to verify that certain major centers solve Problem B, namely those given by trilinears $\sin^p A : \sin^p B : \sin^p C$. As p ranges through the real numbers, these points form the *power curve*. See **[8]**, **[5]**, and Figure 3.

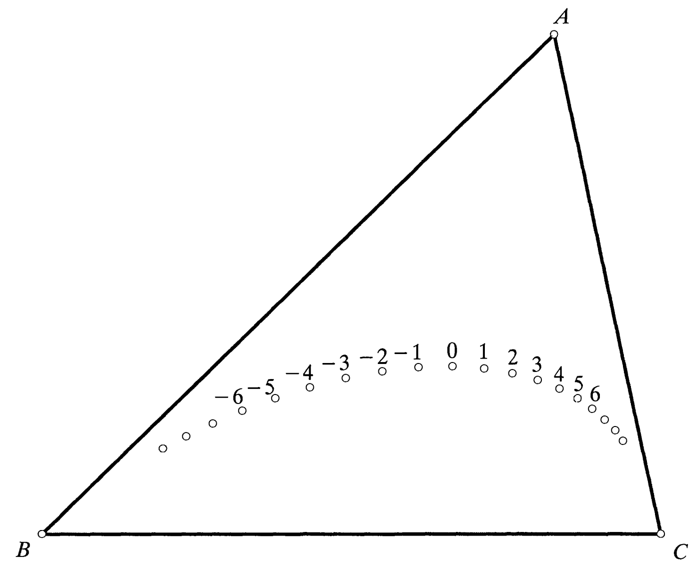


Figure 3. Nineteen points on the power curve, including centroid ($p = -1$), incenter ($p = 0$), and symmedian point ($p = 1$).

Next, suppose in the statement of Problem B that Y is a point (the function kind of point) but not necessarily a center. Would concurrence of the lines AA', BB', CC' for all points X then force Y to be a center? The answer to this question remains to be found.

6. ANOTHER KIND OF CONJUGATE? We return once again to Problem A. This time, we fix $X =$ orthocenter $= \sec A : \sec B : \sec C$, so that

$$M = \begin{pmatrix} \sec A & \sec B & \sec C \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad M^{-1} = \frac{1}{\sec^3 A}\begin{pmatrix} 1 & -\sec B & -\sec C \\ 0 & \sec A & 0 \\ 0 & 0 & \sec A \end{pmatrix}.$$

The transformation $[\alpha : \beta : \gamma]_{ABC} = [\alpha' : \beta' : \gamma']_{XBC} DM$ can be written out as in (4). Here, however, in order to keep $\delta_1 = \delta_2 = \delta_3 = 1$, we shall confine our attention to *acute* triangles only!

It is tedious but not difficult to verify that if $Y = f(A) : f(B) : f(C)$ then the three lines AA', BB', CC' concur if and only if

$$\begin{vmatrix} 0 & -f(C_B)[B] & f(B_C)[C] \\ f(C_A)[A] & 0 & -f(A_C)[C] \\ -f(B_A)[A] & -f(A_B)[B] & 0 \end{vmatrix} = 0, \tag{13}$$

where $[A] = \sqrt{\sec^2 B + \sec^2 C + 2\sec B \sec C \cos A}$, and $[B]$ and $[C]$ are obtained from $[A]$ by the usual cyclic permutations of A, B, C. Obviously, (13) holds, since $C_B = A_B$, $C_A = B_A$, and $A_C = B_C$. See Figure 4. Using (2), the point of concurrence is found to be

$$\begin{aligned} BB' \cap AA' &= f(A_C)f(A_B)[B][C] : f(A_C)f(B_A)[C][A] : f(A_B)f(C_A)[A][B] \\ &= \frac{1}{f(B_A)[A]} : \frac{1}{f(A_B)[B]} : \frac{1}{f(B_C)[C]} \\ &= \frac{\csc 2A}{f(\pi/2 - A)} : \frac{\csc 2B}{f(\pi/2 - B)} : \frac{\csc 2C}{f(\pi/2 - C)}. \end{aligned}$$

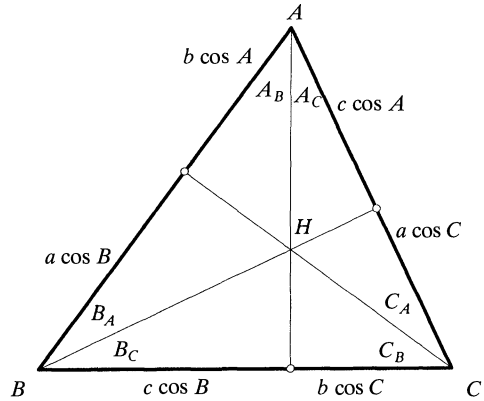


Figure 4. Orthocenter, H, and associated angles and lengths:
$\angle BHC = \pi - A$, $\angle CHA = \pi - B$, $\angle AHB = \pi - C$,
$C_B = A_B$, $C_A = B_A$, $A_C = B_C$.

We write this point of concurrence as $Y^{\perp}$. It is easy to see that for any major center Y, we have $(Y^{\perp})^{\perp} = Y$. Perhaps $Y^{\perp}$ could be called the *orthoconjugate* of Y if someone could find a satisfactory way to extend $(\;)^{\perp}$ to obtuse triangles.

7. CONCLUSION. We wish to emphasize the main point of this paper by returning to Question A. For any given procedure x (i.e., center X), if you attempt to carry out the construction, you are forced to treat Y not as a mere *point*—that's the traditional understanding of a triangle center—but rather as a *function*, here applied to three different triangles. This distinction between point and function really does make a difference, since, as we have seen, Problems A and B have solutions that can be understood only as functions.

REFERENCES

1. Carl B. Boyer, *History of Analytic Geometry*, The Scripta Mathematica Studies, nos. 6 and 7, Yeshiva University, New York, 1956.
2. H. S. M. Coxeter, Some applications of trilinear coordinates, *Linear Algebra Appl.* 226–228 (1995) 375–388.
3. Philip J. Davis, The rise, fall, and possible transfiguration of triangle geometry: a mini-history, *Amer. Math. Monthly* 102 (1995) 204–214.
4. R. H. Eddy and R. Fritsch, The conics of Ludwig Kiepert: a comprehensive lesson in the geometry of the triangle, *Math. Magazine* 67 (1994) 188–205.
5. J. T. Groenman and R. H. Eddy, Problem 858 and Solution, *Crux Mathematicorum* 10 (1984) 306–307.
6. Clark Kimberling, Triangle centers as functions, *Rocky Mtn. J. Math.* 23 (1993) 1269–1286.
7. Clark Kimberling, Hofstadter points, *Nieuw Archief voor Wiskunde* 12 (1994) 109–114.
8. Clark Kimberling, Central points and central lines in the plane of a triangle, *Math. Magazine* 67 (1994) 163–187.
9. Adrian Oldknow, The Euler-Gergonne-Soddy Triangle of a Triangle, *Amer. Math. Monthly* 103 (1996) 319–329.
10. Guido M. Pinkernell, Cubic Curves in the Triangle Plane, *J. Geom.* 55 (1996) 141–161.

Department of Mathematics
University of Evansville
1800 Lincoln Avenue
Evansville, IN 47722
ck6@evansville.edu

On Lambert's proof of the irrationality of π

M. Laczkovich

The irrationality of π was first proved by J. H. Lambert in 1761 in his paper [**3**] (reprinted in [**4**, pp. 112–159]). Lambert's argument is the following. First he proves the formula

$$\tan x = \cfrac{x}{1 - \cfrac{x^2}{3 - \cfrac{x^2}{5 - \cfrac{x^2}{\ddots}}}} \qquad (1)$$

Then Lambert shows, by an argument of infinite descent, that if $x \neq 0$ is rational then the right hand side of (1) is irrational. Since $\tan(\pi/4) = 1$ is rational, this implies that π is irrational.

Lambert's proof is seldom reproduced in books on number theory. The reason is clear: a rigorous proof of (1) cannot avoid questions of convergence of continued fractions, and if our aim is just to prove the irrationality of π then this digression is not worth the effort. (The last monograph that gives Lambert's argument in detail seems to be Chrystal's Algebra [**1**].) The "usual" proofs avoid continued fractions and use variants of Hermite's idea: if π were rational then certain sums or integrals would be integers, contradicting estimates showing that the actual values lie in $(0, 1)$; see Niven's book [**5**]. In the notes on Chapter 2, Niven also gives a list of papers following this line. J. Popken published several papers on the subject that contain variants of Hermite's argument [**6**], [**7**], [**8**]. The paper [**9**] is different; here Popken reproduces Lambert's computation and infers, in a particularly simple way, Lambert's theorem: if $x \neq 0$ is rational then $\tan x$ is irrational.

In this paper we further simplify [**9**] by replacing its computations with Gauss' functional equation. This gives a very simple proof of the irrationality of $\tan x$ (and also of $f(x)$ for a wide class of other functions) whenever $x \neq 0$ is rational. The irrationality of π follows. We also give a self-contained proof of (1) using the same device.

Lambert's original computation leading to (1) was somewhat tedious; he divided the power series of $\sin x$ by that of $\cos x$ using a version of Euclid's algorithm, and determined the quotients and the remainders. This computation was simplified by Gauss [**2**], who determined the continued fraction expansions of the hypergeometric series using their functional equations. If we want to prove only (1), then, following Gauss' argument, we may restrict our attention to the one-parameter family

$$f_k(x) = 1 - \frac{x^2}{k} + \frac{x^4}{k(k+1) \cdot 2!} - \frac{x^6}{k(k+1)(k+2) \cdot 3!} + \cdots.$$

It is easy to see that the series defining f_k converges for every x and for every

$k \neq 0, -1, -2, \ldots$. A simple computation shows that

$$\text{if } k = 1/2 \text{ then } k(k+1)\cdots(k+n-1)\cdot n! = (2n)!/4^n,$$

and

$$\text{if } k = 3/2 \text{ then } k(k+1)\cdots(k+n-1)\cdot n! = (2n+1)!/4^n.$$

Therefore we have

$$f_{1/2}(x) = \cos(2x) \text{ and } f_{3/2}(x) = \frac{\sin(2x)}{2x}$$

for every x. It is also easy to check, by comparing the coefficients of x^{2n}, that

$$\frac{x^2}{k(k+1)} f_{k+2}(x) = f_{k+1}(x) - f_k(x) \qquad (2)$$

for every x and for every $k \neq 0, -1, -2, \ldots$. In the proof of the following theorem we combine (2) with the argument of [**9**].

Theorem 1. *If $x \neq 0$ and x^2 is rational, then $f_k(x) \neq 0$ and $f_{k+1}(x)/f_k(x)$ is irrational for every $k \in \mathbf{Q}$, $k \neq 0, -1, -2, \ldots$.*

Proof: First we show that

$$\lim_{r\to\infty} f_r(x) = 1. \qquad (3)$$

Indeed, since $x^{2n}/n! \to 0$ as $n \to \infty$, there is some $K > 0$ such that $|x^{2n}/n!| \leq K$ for every n. Therefore, if $r > 1$, then $|f_r(x) - 1| \leq \sum_{n=1}^{\infty} K/r^n = K/(r-1)$, from which (3) follows.

Let x be a nonzero real number such that x^2 is rational, let $k \in \mathbf{Q}$, $k \neq 0, -1, -2, \ldots$ be fixed, and suppose that $f_k(x) = 0$ or $f_{k+1}(x)/f_k(x)$ is rational. Then $f_k(x)$ and $f_{k+1}(x)$ are both integer multiples of the same quantity: say $f_k(x) = ay$ and $f_{k+1}(x) = by$ for integers a and b. We allow a or b to be zero. But y cannot be zero, since it would then follow from (2) that $f_{k+n}(x) = 0$ for every $n = 1, 2, \ldots$, which would contradict (3).

Let q be a positive integer such that (bq/k), (kq/x^2), and (q/x^2) are all integers. Now let $G_0 = f_k(x)$ and

$$G_n = \frac{q^n}{k(k+1)\cdots(k+n-1)} f_{k+n}(x) \quad (n = 1, 2, \ldots)$$

for each $n = 1, 2, \ldots$. Then $G_0 = ay$, $G_1 = (bq/k)y$, and from (2) we can calculate that

$$G_{n+2} = \left(\frac{kq}{x^2} + \frac{q}{x^2}n\right)G_{n+1} - \left(\frac{q^2}{x^2}\right)G_n \qquad (4)$$

for every $n = 0, 1, \ldots$. The coefficients in (4) are integers, so G_n is an integer multiple of y for every n. Since $f_{k+n}(x) \to 1$ by (3) and $q^n/[k(k+1)\cdots(k+n-1)] \to 0$, we have $G_n \to 0$. But $f_{k+n}(x) \to 1$ also implies that G_n is positive for all sufficiently large n. Positive integer multiples of y cannot converge to zero. The contradiction means that $f_k(x)$ and $f_{k+1}(x)$ cannot both be integer multiples of the same quantity. ∎

Corollary 2. *π^2 is irrational.*

Proof: $f_{1/2}(\pi/4) = \cos(\pi/2) = 0$. ∎

Corollary 3. *If $x \neq 0$ is rational, then* $\tan x$ *is irrational.*

Proof: Since $(x/2)^2$ is nonzero and rational, $f_{3/2}(x/2)/f_{1/2}(x/2) = (\tan x)/x$ is irrational, and then so is $\tan x$. ∎

Although we eliminated (1) from the proof, for the sake of completeness we give a simple and self-contained proof of (1) using (2). We prove that (1) holds for every complex number x. The continued fraction

$$\cfrac{b_1}{1+\cfrac{b_2}{1+\ddots\cfrac{}{1+\cfrac{b_{n-1}}{1+b_n}}}}$$

will be denoted by $[b_1, \ldots, b_n]$. Since occasionally we may have to divide by zero, we add to the set $\mathbf{C}$ of complex numbers an infinite element ∞, and adopt the following conventions: (i) $a/0 = \infty$ ($a \in \mathbf{C}$, $a \neq 0$); (ii) $a/\infty = 0$, ($a \in \mathbf{C}$); and (iii) $a + \infty = a - \infty = \infty$ ($a \in \mathbf{C}$). It is easy to see, using induction on n, that

$$\text{if } |b_i| \leq 1/4 \text{ for every } i = 1, \ldots, n-1 \text{ and if } |b_n| \leq 1/2, \text{ then } |[b_1, \ldots, b_n]| \leq 1/2. \tag{5}$$

We show next that if $|b_i| \leq 1/4$ for every $i = 1, \ldots, n$ and if $|\delta| \leq 1/4$, then

$$|[b_1, \ldots, b_{n-1}, b_n + \delta] - [b_1, \ldots, b_n]| \leq |\delta|. \tag{6}$$

This is clearly true for $n = 1$. Let $n > 1$, and suppose (6) is true for $n - 1$. Let $|b_i| \leq 1/4$ $(i = 1, \ldots, n)$ and $|\delta| \leq 1/4$. Denoting $u = [b_2, \ldots, b_n]$ and $v = [b_2, \ldots, b_{n-1}, b_n + \delta]$, we have $|u|, |v| \leq 1/2$ by (5), and $|v - u| \leq |\delta|$ by the induction hypothesis. Then

$$\begin{aligned} |[b_1, \ldots, b_{n-1}, b_n + \delta] - [b_1, \ldots, b_n]| &= \left| \frac{b_1}{1+v} - \frac{b_1}{1+u} \right| \\ &= \left| \frac{b_1(v-u)}{(1+u)(1+v)} \right| \\ &\leq \frac{(1/4)|\delta|}{(1-(1/2))(1-(1/2))} = |\delta|, \end{aligned}$$

which completes the proof.

Now let $x \neq 0$ be fixed. Let $k = 1/2$, and put $a_n = f_{n+(3/2)}(x)/f_{n+(1/2)}(x)$ $(n = 0, 1, \ldots)$. By (2) we have

$$a_n = \cfrac{1}{1 - \cfrac{x^2}{(n+(1/2))(n+(3/2))} a_{n+1}} \quad (n = 0, 1, \ldots).$$

Since $a_0 = \tan(2x)/(2x)$, this implies

$$\tan(2x)/(2x) = \left[1, -\frac{x^2}{(1/2)\cdot(3/2)}, \frac{x^2}{(3/2)\cdot(5/2)}, \dots, -\frac{x^2}{(n-(1/2))(n+(1/2))}a_n\right]$$

for every n. Replacing x by $x/2$, and multiplying by x, we obtain

$$\tan x = \left[x, -\frac{x^2}{1\cdot 3}, \frac{x^2}{3\cdot 5}, \dots, -\frac{x^2}{(2n-1)(2n+1)}a_n\right]. \tag{7}$$

Let $N > 1$ be such that $x^2/((2n-1)(2n+1)) < 1/4$ and $a_n \in (0,2)$ for every $n \geq N$ (recall that $\lim_{n\to\infty} a_n = 1$ by (3)). Let

$$P_n = \left[-\frac{x^2}{(2N+1)(2N+3)}, -\frac{x^2}{(2N+3)(2N+5)}, \dots, -\frac{x^2}{(2n-1)(2n+1)}\right]$$

and

$$Q_n = \left[-\frac{x^2}{(2N+1)(2N+3)}, -\frac{x^2}{(2N+3)(2N+5)}, -\frac{x^2}{(2n-1)(2n+1)}a_n\right]$$

for every $n > N$. Then (6) ensures that $|P_n - Q_n| \leq |a_n - 1|$ for every $n > N$. Let

$$F_N(y) = \left[x, -\frac{x^2}{1\cdot 3}, -\frac{x^2}{3\cdot 5}, \dots, -\frac{x^2}{(2N-1)(2N+1)}, y\right].$$

It is easy to check that F_N (as a function of y) is a homeomorphism of $\overline{\mathbf{C}} = \mathbf{C} \cup \{\infty\}$ onto itself; in fact, F_N is a fractional linear transformation. Since $\tan x = F_N(Q_n)$ by (7), we have $Q_n = F_N^{-1}(\tan x)$ for every $n > N$. Since $|P_n - Q_n| \leq |a_n - 1| \to 0$ as $n \to \infty$, this implies $\lim_{n\to\infty} P_n = F_N^{-1}(\tan x)$, and hence, by the continuity of F_N, we obtain $\tan x = \lim_{n\to\infty} F_N(P_n)$. However,

$$F_N(P_n) = \cfrac{x}{1 - \cfrac{x^2}{3 - \cfrac{x^2}{\ddots\, (2n-1) - \cfrac{x^2}{2n+1}}}} \overset{\text{def}}{=} R_n.$$

Since the right hand side of (1) is defined as $\lim_{n\to\infty} R_n$, this proves (1). ■

Note that (1) is valid for every $x \in \mathbf{C}$, even for $x = (\pi/2) + k\pi$, when $\tan x$ is to be interpreted as ∞. The conventions concerning ∞ may be needed for other values of x, too, in order to compute some of the "convergents" R_n. (Take, for example, $x = \sqrt{3}$.) However, for every given x, R_n can be computed using finite numbers only, if n is large enough. Indeed, it is easy to see that there is a finite set S (depending on x) such that for every $y \notin S$, the computation of $F_N(y)$ does not

involve ∞. Since $\lim_{n\to\infty} a_n = 1$, it follows from (2) that $a_n \neq 1$ if $n > n_0$. From this one can prove that $P_n \neq Q_n$ for $n > n_0$, and hence every number occurs in the sequence P_n only a finite number of times. This implies that for $n > n_1$ we have $P_n \notin S$, and then the computation of $R_n = F_N(P_n)$ needs finite numbers only.

REFERENCES

1. G. Chrystal, *Algebra*, Chelsea, New York, 1958.
2. C. F. Gauss, Disquisitiones generales circa seriem infinitam $1 + \frac{\alpha\beta}{1\cdot\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{1\cdot 2\cdot\gamma(\gamma+1)}xx + \frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1\cdot 2\cdot 3\cdot\gamma(\gamma+1)(\gamma+2)}x^3$ + etc, *Comment. soc. reg. sci. Gottingensis rec.* **2**, 1813.
3. J. H. Lambert, Mémoire sur quelques propriétés remarquables des quantités transcendantes circulaires et logarithmiques, *Mémoires de l'Académie des sciences de Berlin* [17] (1761), 1768, 265–322.
4. Iohannis Henrici Lambert, *Opera Mathematica Vol. II*, edited by A. Speiser. Orell Füssli Verlag, Zürich, 1948.
5. I. Niven, *Irrational Numbers*, The Carus Math. Monographs No. 11, Mathematical Association of America, Washington, D.C., 1967.
6. J. Popken, On Lambert's proof for the irrationality of π, *Indagationes Math.* **2** (1940), 279–281.
7. J. Popken, Over de onmeetbaarheid van π, *Euclides* **17** (1941), 217-–227.
8. J. Popken, Over de irrationaliteit van de tangens van een rational getal, *Norsk Matematisk Tidsskrift* **26** (1944), 66–70.
9. J. Popken, Over de irrationaliteit van π, *Rapport Math. Centrum ZW* 1948-014, 1–5.

Department of Analysis
Eötvös Loránd University
Budapest, Múzeum krt. 6-8
Hungary 1088
laczk@cs.elte.hu

...he talked of a learned monk of many centuries ago, who did hit upon a way of multiplying numbers. That in itself I might understand, for it was simple, but the adding of each last two figures to make the next. To wit, one, two, three, five, eight, thirteen, one and twenty, and thus forward as you may will. Mr B. averred that he himself did believe these numbers appeared, though secretly, in many places in nature, as it were a divine cipher that all living things must copy, for that the ratio between its successive numbers was that also of a secret of the Greeks, who did discover a perfect proportion, I believe he said it to be of one to one and six tenths. He pointed to all that chanced about us, and said that these numbers might be read therein; and cited other examples, that I forget now except that many accorded with the order of petals and leaves in trees and herbs, I know not what.

John Fowles, *A Maggot*, New American Library, 1985
Contributed by Irving Adler, North Bennington, VT

NOTES

Edited by **Jimmie D. Lawson**

A Simple Congruence modulo p

Winfried Kohnen

Congruences for prime numbers p have always been of great interest. Examples include Fermat's Little Theorem ($n^p \equiv n \pmod{p}$) or Wilson's theorem ($(p-1)! \equiv -1 \pmod{p}$). In the following we consider the congruence relation modulo p extended to the ring of rational numbers with denominators not divisible by p. For such fractions $m/n \equiv r/s \pmod{p}$ if and only if $ms \equiv nr \pmod{p}$, and the residue class of m/n is the residue class of m times the inverse of the residue class of n in $\mathbf{Z}_p$.

The purpose of this note is to state and prove the following result.

Theorem. *Let p be an odd prime. Then*

$$\sum_{k=1}^{p-1} \frac{1}{k \cdot 2^k} \equiv \sum_{k=1}^{(p-1)/2} \frac{(-1)^{k-1}}{k} \pmod{p}. \tag{1}$$

Proof: First note the identity

$$\sum_{k=1}^{N} \frac{1}{k}(1-X)^k = \sum_{k=1}^{N} \frac{(-1)^k}{k}\binom{N}{k}(X^k - 1) \; (N \in \mathbf{N}, x \in \mathbf{R}). \tag{2}$$

Indeed, the derivative of the left-hand side of (2) is

$$-\sum_{k=1}^{N} (1-X)^{k-1} = -\frac{1-(1-X)^N}{1-(1-X)} = \frac{(1-X)^N - 1}{X},$$

while the derivative of the right-hand side is

$$\sum_{k=1}^{N} (-1)^k \binom{N}{k} X^{k-1}.$$

Hence the derivative of both sides are equal. Also (2) is true for $X = 1$.

In (2) we set $N = p - 1$ and $X = -1$. From $p - k \equiv -k \pmod{p}$, we deduce

$$\binom{p-1}{k} = \frac{(p-1)\cdots(p-k)}{k!} \equiv (-1)^k \pmod{p}$$

and

$$\sum_{k=1}^{p-1} \frac{1}{k} = \sum_{k=1}^{p-1} \frac{1}{p-k} \equiv -\sum_{k=1}^{p-1} \frac{1}{k} \equiv 0 \pmod{p},$$

and thus equation (2) simplifies to

$$\sum_{k=1}^{p-1} \frac{2^k}{k} \equiv \sum_{k=1}^{p-1} \frac{(-1)^k}{k} \pmod p. \tag{3}$$

In the sum on the left we replace k by $p - k \equiv -k \pmod p$ and use Fermat's Little Theorem to obtain

$$\sum_{k=1}^{p-1} \frac{2^k}{k} \equiv -2^p \sum_{k=1}^{p-1} \frac{1}{k \cdot 2^k} \equiv -2 \sum_{k=1}^{p-1} \frac{1}{k \cdot 2^k} \pmod p.$$

The sum on the right of (3) we rewrite as

$$\sum_{k=1}^{(p-1)/2} \frac{(-1)^k}{k} + \sum_{k=1}^{(p-1)/2} \frac{(-1)^{p-k}}{p-k} \equiv 2 \sum_{k=1}^{(p-1)/2} \frac{(-1)^k}{k} \pmod p.$$

This proves (1).

In the literature, congruences of a type similar to (1) can be found; however, in general they are of a much deeper nature. For example, in [1] with the help of properties of the Pell sequence $((1+\sqrt{2})^n)_{n \in N}$ it is shown that

$$\sum_{k=1}^{(p-1)/2} \frac{1}{k \cdot 2^k} \equiv \sum_{k=1}^{[3p/4]} \frac{(-1)^{k-1}}{k} \pmod p. \tag{4}$$

It seems unlikely that (4) can be proved with the simple approach we have used here.

REFERENCE

1. Zhi-Wei Sun, A congruence for primes, *Proc. Amer. Math. Soc.* 123 (1995), 1341–1346.

Mathematisches Institut
Universität Heidelberg
Im Neuenheimer Feld 288
69120 Heidelberg, Germany
winfried@mathi.uni-heidelberg.de

A Geometrical Method for Finding an Explicit Formula for the Greatest Common Divisor

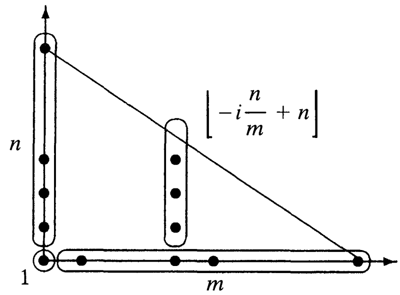
Marcelo Polezzi

This note presents an explicit formula for the greatest common divisor (g.c.d.) of two integers derived using a simple geometrical argument.

In [1], chapter 3, an expression was deduced, from which one can easily obtain a formula for the g.c.d. as a particular case. However, the derivation of that expression is very tiring and lengthy.

Here is the result to be proved:

Theorem. *Let m and n be positive integers. Then* $g.c.d.(m,n) = 2\sum_{i=1}^{m-1}\left\lfloor i\frac{n}{m}\right\rfloor + (m+n) - mn$, *where* $\lfloor x \rfloor$ *is the greatest integer less than or equal to* x.



Proof: Consider the triangle $M = \left\{(x,y) \in \mathbb{R}^2 | x \geq 0;\ y \geq 0;\ y \leq -\frac{n}{m}x + n\right\}$. We have

$$\#(M \cap \mathbb{Z}^2) = \sum_{i=1}^{m-1}\left\lfloor -i\frac{n}{m} + n\right\rfloor + (m+n+1)$$
$$= \sum_{i=1}^{m-1}\left\lfloor (m-i)\frac{n}{m}\right\rfloor + (m+n+1) = \sum_{i=1}^{m-1}\left\lfloor i\frac{n}{m}\right\rfloor + (m+n+1). \quad (1)$$

On the other hand, by considering the triangle as half of a rectangle, we obtain

$$\#(M \cap \mathbb{Z}^2) = \frac{(m+1)(n+1) + (d+1)}{2} \quad (2)$$

where $d = g.c.d.(m,n)$, since the number of lattice points on the hypotenuse is equal to $(d+1)$. In fact, let $y_i = -(n/m)x_i + n$. The set of integers x_i between 0 and m such that y_i is an integer is

$$\left\{0, \frac{m}{d}, 2\frac{m}{d}, \ldots, (d-1)\frac{m}{d}, m\right\}.$$

Hence, equating (1) and (2), we have

$$\sum_{i=1}^{m-1}\left\lfloor i\frac{n}{m}\right\rfloor + (m+n+1) = \frac{(m+1)(n+1) + (d+1)}{2}.$$

Therefore, $d = 2\sum_{i=1}^{m-1}\left\lfloor i\frac{n}{m}\right\rfloor + (m+n) - mn$.

Remark. The formula clearly holds for $n = 0$.

REFERENCE

1. R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics—A Foundation for Computer Science*, Addison-Wesley Publishing Company, 1994.

Departamento de Matemática
Instituto de Biociências, Letras e Ciências Exatas
Universidade Estadual Paulista
São José do Rio Preto-S.P-CP 136
Brazil

THE EVOLUTION OF . . .

Edited by **Abe Shenitzer**
Mathematics, York University, North York, Ontario M3J 1P3, Canada

On the Historical Development of Infinitesimal Mathematics

translated by Abe Shenitzer with the editorial assistance of Hardy Grant

Detlef Laugwitz

PART I. THE ALGORITHMIC THINKING OF LEIBNIZ AND EULER

1. LEIBNIZ AND L'HOSPITAL. Exactly 300 years ago there appeared in Paris the first book on differential calculus, *Analyse des infiniment petits*, by Marquis Guillaume F.A. L'Hospital (1661–1704). The book was based on materials supplied by Johann Bernoulli (1667–1748). Johann and his older brother Jakob (1654–1705) were the first "comrades-in-arms" of Gottfried Wilhelm Leibniz (1646–1716), one of the two discoverers of the calculus. In 1684 Leibniz published for the first time a few simple applications of the differential calculus, without making any attempt to provide clear justifications for them. But we find such attempts in his letters to judicious contemporaries. For example, on 30 March 1690, Leibniz wrote to John Wallis (1616–1703): "It is useful to regard quantities as infinitely small, so that, when their ratio is sought, they are omitted, rather than viewed as 0, when they turn up next to quantities that are incomparably larger. If we have $x + dx$, then dx is omitted. Similarly, we cannot let $x\,dx$ and $(dx)^2$ stand next to one another. Thus if we have to differentiate xy, we write $(x + dx)(y + dy) - xy = x\,dy + y\,dx + dx\,dy$. But here $dx\,dy$ should be omitted, for it is incomparably smaller than $x\,dy + y\,dx$. Hence, in each particular case, the error is smaller than any finite quantity." (Leibniz, Math. Schriften IV, 63; our translation. D.L., A.S.)

Leibniz proceeds pragmatically. He states a rule which gives a correct result in 'every special case.' At first sight, it all looks like facts based on experience, but theoretical justifications are in the offing. Leibniz seems to add to the system of quantities apparently measured in terms of real numbers ideal elements $dx, dy, \ldots$ such that a (positive) dx is smaller than every (positive) real quantity. Then the algebraic rules of computation are applied in the usual manner. The worked examples show that at the end of a computation the differentials are replaced by 0s. (In the last step we are, to some extent, anticipating history, for it was soon noticed that the stipulation pertaining to the omitting of a quantity against an incomparably greater one can be problematic. A case in point is $(x + dx)^2 - x^2 = 2x \cdot dx + (dx)^2$ for $x = 0$!).

This recipe can be used for algebraic expressions. As early as 1684 Leibniz obtained the root formula

$$d\sqrt{y} = \sqrt{y + dy} - \sqrt{y} = \frac{dy}{\sqrt{y + dy} + \sqrt{y}},$$

which reduces to $d\sqrt{y} = dy/2\sqrt{y}$. The latter is justified because dy may be omitted against y.

At the time when Bernoulli and L'Hospital were working on the latter's book, the calculus could already boast of significant achievements in geometry and in mechanics, as well as in variational problems. The latter is certainly not beginners' stuff. A method for the solution of problems had been found, a method of analysis in the original sense of the word. The subsequent designation of the differential and integral calculus as "analysis" is derived from the title of the first relevant book. L'Hospital tried at first to present the subject synthetically, i.e., to begin with a statement of principles (axioms and postulates) followed by deductions. What Leibniz expressed in words was postulated in the book in the form of the equation $y + dy = y$. This produced shock, for in a computation dy was not supposed to be initially put equal to 0.

As a philosopher, Leibniz gave a great deal of thought to equality and equivalence relations, but there is no trace of this in his calculus. Now, in nonstandard analysis, we write $a = b$ for equality proper and $\approx$ for "almost-equality," i.e., we write $a \approx b$ if $a - b$ is infinitely small. Moreover, we carefully differentiate $\mathbb{R}$, the system of real numbers, from the larger system that includes, in addition, the infinitely small numbers. In our set-based symbolism, which is just as ahistorical as the symbol $\approx$, we denote the latter system by ${}^*\mathbb{R}$ and speak of the hyperreal numbers. A genuine historian may be horrified, but so long as we are aware that we are using our richer professional language, and that in some cases this language does not quite reflect the historical state of affairs, not much harm is done. Every historian must face up to the fact that the language of the sources is less, or differently, expressive than the corresponding modern professional language.

Using our language, it is possible to state the recipe of Leibniz and L'Hospital very clearly: We compute in ${}^*\mathbb{R}$. If at the end of a calculation there appears an almost-equality $a \approx b$ and a is almost-equal to a real number, then b must also be almost-equal to this real number. Those who want to be really fancy, can speak of a homomorphism st (= standard part of) from a part of ${}^*\mathbb{R}$ to $\mathbb{R}$. The reader can easily verify that $st(a + b) = st\, a + st\, b$, $st(a \cdot b) = (st\, a)(st\, b)$, provided that all standard parts exist. That this is not always the case is shown by the example of $1/dx$.

There is something else we are about to modernize. From Leibniz's letter to Wallis we know Leibniz's proof of the product formula: He writes an equation in terms of differentials, and if we go over in it to standard parts, then we obtain $0 = 0$ and are not much wiser. This is why we prefer to use quotients of differentials (this happened soon after) and view x and y in the derivation as functions of t. Then

$$dx = x(t + dt) - x(t),\ etc.,$$

and

$$\frac{d(xy)}{dt} = \frac{1}{dt}[(x + dx)(y + dy) - xy] = x\frac{dy}{dt} + \frac{dx}{dt}y + \frac{dx}{dt}dy.$$

The derivative function was accepted only after 1800. We use it immediately: If for a fixed $t = t_0$ all differential quotients dx/dt have the same standard part regardless of the choice of the infinitely small dt, then this standard part is called the derivative $x'(t_0)$; the function $x(t)$ is said to be differentiable at t_0. After going over to standard parts, we read off from our formula the following result: If x and y are differentiable, then so is $x \cdot y$ and we have the product formula $(xy)' = xy' + x'y$.

Unfortunately, since Leibniz's time everything has become a bit more complicated, but this is the price we must pay for precision. In general, the differential quotient is a hyperreal number, while the derivative is always real for real numbers. We leave it to the reader to prove the quotient formula and the chain rule. By now we have justified the most important part of the differential calculus. For the time being, we leave over integration and turn to Euler and his 'algebraic analysis.'

2. LEONHARD EULER AND THE ANALYSIS OF THE INFINITE. Euler was born in 1707 in Basel. He learned the newest mathematics from Johann Bernoulli. This was as natural for him as learning the multiplication table is for others. At the age of 20 he left Basel never to return. He worked in St. Petersburg, from 1741 to 1766 in Berlin, and until the end of his life in 1783 again in St. Petersburg. He was the greatest mathematician of the 18th century in all areas, including applications. The recent little book by E. Fellmann, head of the Basel Euler-Archiv, is a must. The book that appeared on the 200th anniversary of Euler's death is also valuable (Fellmann 1995, Euler 1983). We are about to present small excerpts from Euler's monumental work, chosen to illustrate the manner of thinking that prevailed in his time.

Leibniz and the Bernoullis used infinitely small numbers in the differential calculus, but they also made occasional use of infinitely large numbers. It was quite natural for Euler to work with infinite natural numbers and with infinite series including summands with infinite index. His paper on harmonic series (*De progressionibus harmonicis observationes*, Euler Op. ser. I vol. 14, 73–86), written in 1734, is highly recommended. It was common knowledge that integration of the hyperbolic function $y = 1/x$ yields the natural logarithm. If we form upper and lower sums that extend to integral division points then, like Euler, we obtain

$$\sum_{n=1}^{N} \frac{1}{n} = \log N + C_N,$$

and for infinitely large N we always have $C_N \approx C = 0.577\ldots$, where C is the so-called Euler constant. This implies that

$$\begin{aligned}
\sum_{n=1}^{2N} \frac{(-1)^{n+1}}{n} &= 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + - \cdots + \frac{1}{2N-1} - \frac{1}{2N} \\
&= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{2N-1} + \frac{1}{2N} \\
&\quad 2\left(\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2N}\right) \\
&= \sum_{n=1}^{2N} \frac{1}{n} - \sum_{n=1}^{N} \frac{1}{n} = \log 2N + C_{2N} - \log N - C_N \\
&\approx \log 2N - \log N = \log \frac{2N}{N} = \log 2.
\end{aligned}$$

In much the same way Euler evaluated certain previously unknown sums. We suggest that the reader should try to obtain Euler's series for $\log 3 = \log 3N - \log N$.

It is not difficult to verify 'rigorously' these sums of series by considering finite partial sums and passing to the limit. But for Euler there were no limit considerations, and what we want to know is the kind of reasoning employed by him and by his contemporaries. This being so, it would be counterproductive to handle these problems using perfected later means.

According to a view that stubbornly persists in the secondary literature, Euler gave no thought to convergence. And yet his early work devoted to harmonic series contains a convergence criterion; more precisely, a condition for a series to have a finite sum. When a series has a finite sum, what is added after an infinite term is actually infinitely small. Conversely, if what is obtained from the continuation beyond the most infinite term were finite, then the sum of the series would necessarily be infinite. (The last condition holds only for series with nonnegative terms.) Stated as a formula, this criterion asserts that Σa_n is finite (the series converges) if and only if for N infinitely large we always have $\Sigma_{n \geq N} a_n \approx 0$.

Euler was never particularly interested in general conceptual arguments but, as we see here, they were not foreign to him. He was far more interested in algorithmic applications. In the case of the general harmonic series Σn^{-p}, for $0 < p < 1$, he finds that

$$\sum_{n=1}^{2N} \frac{1}{n^p} > N \cdot \frac{1}{(2N)^p} = \frac{1}{2^p} N^{1-p}.$$

Since this is infinitely large, the series diverges. For $p > 1$ Euler concludes that the series converges. All this had been known for a long time on the basis of, say, an appropriate integral test. If Euler re-proves these facts, he does so, presumably, to make use of his convergence criterion which, unlike the integral test, can also be applied to alternating series. For further details see my book of 1986.

Euler's original papers were not easily accessible. They appeared in obscure publications, mostly in Latin, and on rare occasions in French. Hence what influenced larger circles of readers was his textbooks, which soon appeared in German translations provided with (consistently bad) commentaries. It seems that already at that time not all the interested readers had a sufficient knowledge of Latin! For a long time the image of Euler as a mathematician was a reflection of these elementary books rather than of his deeper original works. We must keep this in kind, now that we are about to turn to his *Introductio in analysin infinitorum* (published in 1748). Today, H. Maser's modern translation is readily available as a reprint and includes a usable commentary.

The literal translation of the Latin title is, of course, *An Introduction into the Analysis of the Infinite*, but a more fitting title would be *An Introduction to the Solution of Problems by Means of Infinite* [numbers and processes]. Euler algebraized infinite series, products, and continued fractions. He tested the possibilities resulting from the formal extension of algebraic calculations with finite formulas to calculations with infinite ones. In particular, he treated power series as polynomials of infinite degree. His approach to the exponential function was typical. He defined (!) it as

$$e^x = \left(1 + \frac{x}{N}\right)^N$$

with N an infinite natural number. He made use of the binomial formula and

obtained

$$1 + N \cdot \frac{x}{N} + \frac{1}{2!}\frac{N(N-1)}{N^2}x^2 + \frac{1}{3!}\frac{N(N-1)(N-2)}{N^3}x^3 + \cdots$$
$$= 1 + x + \left(1 - \frac{1}{N}\right)\frac{x^2}{2!} + \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\frac{x^3}{3!} + \cdots$$

At this point things get precarious. Presumably, since $1/N = 2/N = \cdots = 0$, we obtain the power series expansion

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Euler was well aware, and said it elsewhere, that one couldn't simply claim that the sum of infinitely many small errors is 0. After all, an integral is the sum of infinitely many infinitely small rectangular areas. Euler could easily prove the convergence of this series for real and even for complex values by means of his convergence criterion (which he doesn't mention in the book). All that is needed is to majorize the remainder of the series at infinity by means of a geometric progression. Its partial sums of finite length are ultimately arbitrarily close to its sum. Since the partial sums of the same length in the binomial expansion differ from them by arbitrarily little, the asserted convergence follows. In any case, one should write $e^x \approx (1 + x/N)^N$. Incidentally, the argument shows that there is no need to choose a special infinitely large natural number N.

If $y = e^x$, then $x = \log y$. But $y = (1 + x/N)^N$ implies that $x = N(y^{1/N} - 1)$. Since, in general, an N-th root has N complex values, one would expect the logarithm to be infinitely many-valued. This is how, in the algebraic style of his *Introductio*, Euler settled in 1749 the dispute that animated his two great predecessors in the years 1712–1713. Their exchange of letters was published in 1745. The title of Euler's essay is *De la controverse entre MM. Leibniz et Bernoulli sur les logarithmes des nombres negatifs et imaginaires*. (Euler Op. ser. I, vol. 17.)

The essay is an expository masterpiece. It can be read by university and even high school students. It was written in the language of the Berlin court and seems to have been intended for a wide readership. This is indicated by the thorough discussion of details.

Euler comments on the arguments of the two opponents. Bernoulli thought that $\log(-1) = 0$. This can be justified on the basis of the functional equation: $0 = \log 1 = \log(-1)^2 = 2\log(-1)$. Leibniz argued that since the exponential function is positive for all real inputs, the logarithm of -1 cannot be real. Both adduced further justifications for their respective positions. The issue remained unresolved until Euler produced his surprising solution and actually computed explicitly the infinitely many values of the logarithm. For example, take $y = 1$. By the de Moivre formula we have

$$y^{1/N} = \cos\frac{2\pi g}{N} + i \cdot \sin\frac{2\pi g}{N}$$

for every integer g. For a finite g we have $N\ (\cos(2\pi g/N) - 1) \approx 0$ and $N \sin(2\pi g/N) \approx 2\pi g$, so that $\log 1 = 2\pi g$ for every integer g. More generally, for every complex $z \neq 0$ Euler obtained the values $\log z = \log r + i(\alpha + 2\pi g)$, where $z = r(\cos\alpha + i\sin\alpha)$.

The knowledgeable reader will notice that Euler could have obtained the same result more simply by using his famous formula $e^{i\alpha} = \cos\alpha + i \cdot \sin\alpha$ and applying the functional equation of the logarithm to $z = r \cdot e^{i\alpha} = r \cdot e^{i(\alpha + 2\pi g)}$. He had discovered this formula a few years earlier, but it was not yet universally accepted. And not even his algebraic solution of the controversy convinced all his contemporaries; d'Alembert defended Bernoulli's position as late as 1761.

What made this question so important? Why did the greatest minds persistently return to it? A key problem at the beginning of the century was indefinite integration. In the case of differentiation one had an algorithm which when applied to, say, rational functions again yielded rational functions. But the inverse operation, integration, was not so simple. In 1702 Leibniz and Johann Bernoulli independently investigated the integration of rational functions through decomposition into partial fractions. For example,

$$\frac{1}{x^2 - a^2} = \frac{1}{2a}\left\{\frac{1}{x-a} - \frac{1}{x+a}\right\}$$

led to

$$\int \frac{dx}{x^2 - a^2} = \frac{1}{2a} \log \frac{x-a}{x+a}.$$

It made sense to try to use this for $a = i = \sqrt{-1}$, in spite of the fact that it was known that the arctan entered at this point. But this too could be reduced to logarithms of complex numbers. Hence the great interest in logarithms.

One would have come closer to the solution if one had used suitable limits of integration in

$$\int \frac{dx}{x^2 + 1} = \frac{1}{2i} \log \frac{x-i}{x+i}$$

and introduced known values of the arctan function:

$$\pi = \int_{-\infty}^{+\infty} \frac{dx}{x^2+1} = \frac{1}{2i}[\log 1 - \log 1],$$

$$\frac{\pi}{2} = \int_{0}^{+\infty} \frac{dx}{x^2+1} = \frac{1}{2i}[\log 1 - \log(-1)],$$

$$\frac{\pi}{2} = \int_{-\infty}^{0} \frac{dx}{x^2+1} = \frac{1}{2i}[\log(-1) - \log 1].$$

Already the first of these relations shows that 0 is not the only value that can be assigned to $\log 1$, and the two other equations suggest $\log(-1) = \pm i\pi$. Euler's algebraic solution could not answer the question of which value to choose in a particular case. For this one had to wait another century, when Riemann introduced his surfaces which made the handling of many-valued functions transparent. It is worth noting that it was the problem of integration of algebraic (and in particular rational) functions that gave rise to Riemann surfaces.

When it comes to investigations in the *Introductio* connected with other elementary functions, we mention the power-series expansion of the logarithm function, related to the solution of $y = (1 + x/N)^N$ for x. Putting $y = 1 + h$, $x = \log(1 + h)$ we obtain

$$\log(1+h) = (1+h)^{1/N} - 1.$$

After Newton, it was legitimate to use the binomial series for arbitrary real exponents. Hence, for $|h| < 1$, the series:

$$\log(1+h) = N \cdot \frac{1}{N}h + \frac{N}{2} \cdot \frac{1}{N}\left(\frac{1}{N} - 1\right)h^2 + \frac{N}{3!} \cdot \frac{1}{N}\left(\frac{1}{N} - 1\right)\left(\frac{1}{N} - 2\right)h^3$$

$$+ \cdots \approx h - \frac{h^2}{2} + \frac{h^3}{3} - + \cdots$$

3. CRITICISM AND EARLY ATTEMPTS AT JUSTIFICATION. The Achilles heel of Leibniz's calculus was the "equality" $x + dx = x$. The most vehement criticism of the inadequate justification of the calculus came from Paris, the contemporary center of research that had provided Leibniz with powerful stimuli, derived from contacts with Huygens and the study of the writings of Pascal, during his stay from 1672 to 1676. On 2 February 1702 Leibniz wrote a detailed letter to Varignon, his Paris ally, which the latter published immediately. A few years later the Parisians became converts to the new calculus, and after that there were virtually no objections to it on the European continent.

Leibniz provided many justifications. He argued that, much like imaginary numbers, infinite and infinitely small distances can be used without reservations as ideal concepts, without having to be regarded as real objects. It is more convenient to introduce once and for all the concept of the incomparably small instead of always having to speak about magnitudes capable of unlimited decrease. Nevertheless, as is shown by the letter to Wallis quoted earlier, it is this very turn of phrase that justifies the calculus, by showing that the error is invariably less than any assignable magnitude.

This is, so to say, a pragmatic argument: The calculus of differentials is more convenient than talking about constantly decreasing magnitudes, more convenient than the method of limits advocated by Newton. Then comes a 'metaphysical' argument, an appeal to the general continuity principle. Leibniz asserts that the rules of the finite hold for the infinite and conversely. The immediate question is: Which rules? Surely the algebraic rules, which Leibniz carries over without hesitation from the real numbers to differentials. But that is not all. Euler's use of infinitely large natural numbers also accords with the principle of permanence of the rules of operation. Leibniz wrote too little, and what he wrote includes no systematic account. Hence it is impossible to draw unassailable conclusions, based on his usage, concerning the meaning of general pronouncements on the validity of rules. This is different in the case of Euler.

Speaking of Euler. While *his* use of the infinitely large and small in the *Introductio* and in his papers on series *can* be made to harmonize with the principle of permanence of rules of operation, a similar reconciliation of usage and theory is well nigh impossible when it comes to the chapter *On the infinite and on the infinitely small* in his book on the differential calculus (1755). This chapter has generated more confusion than clarity. I won't go into this matter here and refer the interested reader to pp. 206–211 of my book of 1986.

The persuasive power of the calculus derived from the incontestable successes it amassed in the 18th century. But the need for a clear justification was imperative. In this connection we must mention, first of all, J. L. Lagrange, who took over Euler's position in Berlin in 1766. In 1784, one year after Euler's death, he announced a prize competition of the Berlin Academy. He alluded to the fact that many philosophers and mathematicians view the notion of the infinite as inconsis-

tent, and posed the problem of explaining how it is possible to derive so many true results from an inconsistent assumption. In addition, participants were to provide a genuinely mathematical principle that would replace the notion of the infinite without, however, leading to a loss of simplicity and clarity in the derived results.

None of the approximately two dozen responses satisfied Lagrange, so that in 1797, while in Paris, he published his own proposed solution. The long title of his book states clearly what he had in mind: A theory of analytic functions, including the foundations of the differential calculus, free of all consideration of the infinitely small, of the vanishing, of limits and fluxions, returned to the analysis of finite magnitudes.

Lagrange's analytic functions are (formal) power series

$$f(x+h) = f(x) + h \cdot g(x) + h^2[\cdots].$$

He wanted to work with them without involving convergence arguments. His attempt failed, but it has had an abiding influence. The $g(x)$ in the preceding formula stands for the derivative, which he denoted by the symbol $f'(x)$. Since then, the derivative has occupied a foreground position, next to the differential quotient. Since Lagrange also investigated properties of linear order, he gave the first remainder estimates. We add that when it came to applications, he too thought in pragmatic terms. In his analytical mechanics he computed with differentials.

4. NEW DEMANDS FROM PHYSICS. Around 1750 it became clear that the algebraic-algorithmic notion of functions was inadequate from the viewpoint of physical applications. The famous example is that of the vibrating string. Daniel Bernoulli (1700–1782), son of Johann, found that an adequate mathematical treatment of this phenomenon required the use of series such as $\Sigma b_n \sin nx$. He and Euler realized that such series represented fairly arbitrary functions, whose behavior was very different from that of hitherto admissible functions. In particular, they did not have to obey the same "law" throughout their respective domains of definition. Thus

$$\sin x + \frac{\sin 2x}{2} + \frac{\sin 3x}{3} + \cdots$$

is equal to $(\pi - x)/2$ for $0 < x < 2\pi$ and $(3\pi - x)/2$ for $2\pi < x < 4\pi$. The sum function is piecewise linear and has a jump discontinuity at all integral multiples of 2π. Euler obtained this result by summing a geometric progression with ratio $q = re^{ix}$:

$$\sum_{n=0}^{\infty} r^n e^{ix} = \frac{1}{1 - re^{ix}} = \frac{1 - re^{-ix}}{(1 - re^{ix})(1 - re^{-ix})}$$
$$= \frac{1 - r\cos x + ir\sin x}{1 - 2r\cos x + r^2}.$$

Its real part is

$$1 + r\cos x + r^2\cos 2x + r^3\cos 3x + \cdots = \frac{1 - r\cos x}{1 - 2r\cos x + r^2}.$$

This is valid for $0 < r < 1$. Nevertheless, we follow Euler and use this formula

uncritically for $r = 1$. The result is

$$1 + \cos x + \cos 2x + \cos 3x + \cdots = \frac{1}{2}.$$

(1754; see Opera ser. I, 14, pp. 542–584 and 15, pp. 435–497.) Termwise integration of the latter series yields

$$\sum_{n=1}^{\infty} \frac{\sin nx}{n} = \frac{\pi - x}{2} \text{ for } 0 < x < 2\pi.$$

The value of the constant of integration is obtained by putting $x = \pi$ in the series.

What Euler does here is different from what he did 20 years earlier. He no longer computes with possibly infinite values of a divergent series but assigns to it the appropriate numerical value of the expression that gave rise to it. In the case of our geometric progression, the expression in question is $(1 - q)^{-1}$. In three notes written between 1771 and 1773, the physicist Daniel Bernoulli adopted this algorithmic approach and developed it further (D. Bernoulli, Werke Bd. 2, Basel 1982; the editor regards this as an application of the Leibniz principle of continuity).

We may view this method as highly questionable, but its successes were beyond doubt. Just as in the case of differentials, we are challenged to provide a clear justification for a successful *modus operandi.* I provide such a modern justification in my book of 1986 (see pp. 181 ff.).

The modern field-theoretic conception of physics began with the theory of heat developed after 1807 by J. B. Joseph Fourier. This resulted in a viewpoint that regarded all physical magnitudes as solution functions of (partial) differential equations. Since these functions were supposed to correspond to real measurements of physical magnitudes, it followed that the only useful functions were those whose values for real inputs were real. As a result, in the first instance, functions given by formal expressions, including divergent series, were marginalized.

Ahornweg 23
D-64367 Mühltal
Germany

Let $Ax = b$ where A is n-by-n and nonsingular. Denote the columns of A by $a_1, \ldots, a_n$ and those of the identity matrix by $e_1, \ldots, e_n$. Then

$$\begin{aligned}
&\det[a_1, \ldots, a_{i-1}, b, a_{i+1}, \ldots, a_n]/\det A \\
&= \det A^{-1}[a_1, \ldots, a_{i-1}, b, a_{i+1}, \ldots, a_n] \\
&= \det[A^{-1}a_1, \ldots, A^{-1}a_{i-1}, A^{-1}b, A^{-1}a_{i+1}, \ldots, A^{-1}a_n] \\
&= \det[e_1, \ldots, e_{i-1}, x, e_{i+1}, \ldots, e_n] \\
&= x_i, \quad \text{the } i\text{th component of } x.
\end{aligned}$$

Contributed by Kong-Ming Chong, University of Malaya

PROBLEMS AND SOLUTIONS

Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**

with the collaboration of Paul T. Bateman, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Underwood Dudley, Michael A. Filaseta, Ira M. Gessel, Bart Goddard, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Robert Israel, Murray S. Klamkin, Daniel J. Kleitman, Fred Kochman, Frederick W. Luttmann, Frank B. Miles, M. J. Pelling, Richard Pfiefer, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before October 31, 1997; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.*

PROBLEMS

10585. *Proposed by Roger Pinkham, Hoboken, NJ.* Three points are selected independently and at random in a disk of radius one. What is the average distance of the third from the line determined by the first two?

10586. *Proposed by Donald E. Knuth, Stanford University, Stanford, CA.* A certain matrix has m rows and $n = 1 + k^2$ columns. All entries of the matrix are ± 1, and the dot product of any two columns is less than or equal to 0. Prove that the total number of positive entries in the matrix is at most $\frac{1}{2}m(n+k)$, and construct a matrix that achieves this upper bound.

10587. *Proposed by Dave Witte, Oklahoma State University, Stillwater, OK.* Suppose a particle moves continuously in the plane in such a way that the distance between its position at any two different times depends only on the difference between the times. Prove that the particle travels either on a circular arc or on a straight line.

10588. *Proposed by Herbert S. Wilf, University of Pennsylvania, Philadelphia, PA.* Show that

$$\prod_{j\geq 1} e^{-1/j}\left(1+\frac{1}{j}+\frac{1}{2j^2}\right) = \frac{e^{\pi/2}+e^{-\pi/2}}{\pi e^{\gamma}},$$

where γ is Euler's constant.

10589. *Proposed by Paul Bateman, University of Illinois, Urbana, IL and David Bradley, Simon Fraser University, Burnaby, BC, Canada.*

(a) Prove the identity

$$\sum_{j=0}^{2^{k-1}-1} (-1)^{k-1-\eta(j)}(y+j)^k = k!\cdot 2^{(k-1)(k-2)/2}\left(y+(2^{k-1}-1)/2\right),$$

where $\eta(j)$ is the number of ones in the binary expansion of the nonnegative integer j.

(b) Use part **(a)** to infer that there is a positive integer $s = s(k)$ such that every integer n is expressible in the form $n = \epsilon_1 x_1^k + \epsilon_2 x_2^k + \cdots + \epsilon_s x_s^k$ in infinitely many ways, where $\epsilon_i = \pm 1$ for $1 \leq i \leq s$ and where $x_1, x_2, \ldots, x_s$ are distinct positive integers.

10590. *Proposed by David Cox, Amherst College, Amherst, MA.* Fix an integer $n \geq 2$ and let $d_1, d_2, \ldots, d_n$ be positive integers with no common divisor greater than 1. Suppose that d_i divides $d_1 + \cdots + d_n$ for all $i = 1, \ldots, n$.
(**a**) Prove that $d_1 d_2 \cdots d_n$ divides $(d_1 + \cdots + d_n)^{n-2}$.
(**b**) For each $n \geq 3$, give an example to show that the exponent in part (**a**) cannot be made smaller.

10591*. *Proposed by Jeffrey C. Lagarias, AT&T Research, and Thomas J. Richardson, Bell Laboratories, Murray Hill, NJ.* Let F_1, F_2, F_3, F_4 denote the faces of a tetrahedron. For $i = 1, 2, 3, 4$, let α_i denote the solid angle of the vertex opposite face F_i, where the measure of a solid angle is normalized so that a full solid angle is 1, and let β_i denote the area of F_i, where the unit of area is normalized so that the tetrahedron has surface area 1.
(**a**) Prove that $\beta_i \geq \alpha_i$.
(**b**) Generalize to n dimensions.

SOLUTIONS

Subcollections with Large Intersection

10383 [1994, 473]. *Proposed by Kevin Ford (student), University of Illinois, Urbana, IL.* Let $B_1, \ldots, B_s$ denote subsets of a finite set B, and let $\lambda_i = \#(B_i)/\#(B)$ and $\lambda = \lambda_1 + \cdots + \lambda_s$. Show that, for every integer t satisfying $1 \leq t \leq \lambda$, there exist $r_1, \ldots, r_t$ with $r_1 < \cdots < r_t$ and

$$\#(B_{r_1} \cap \cdots \cap B_{r_t}) \geq (\lambda - t + 1)\binom{s}{t}^{-1} \#(B).$$

Solution by Mark Bowron, Houston, TX. It suffices to show that the quantity on the right is the average size of the intersection over all $\binom{s}{t}$ choices of $\{r_i\}$; in particular, $N \geq (\lambda - t + 1)\#(B)$, where N is the sum of the intersection sizes. With $B = \{x_1, \ldots, x_n\}$, let u_k be the number of sequences $\{r_i\}$ such that $x_k \in \bigcap_{i=1}^{t} B_{r_i}$, and let v_k be the number of members of $\{B_i\}$ that contain x_k. Then

$$N = \sum_{k=1}^{n} u_k = \sum_{k=1}^{n} \binom{v_k}{t} \geq \sum_{k=1}^{n} (v_k - t + 1) = \sum_{i=1}^{s} \#(B_i) - nt + n = (\lambda - t + 1)\#(B).$$

Editorial comment. The proposer and Víctor Hernández observed that the result holds also for events $B_1, \ldots, B_s$ in a probability space, where $\lambda_i = P(B_i)$ and we seek a lower bound on $\max P\left(\bigcap B_{r_i}\right)$. Some solvers observed that the bound holds also when $\lambda \leq t \leq \lambda + 1$. A. N. 't Woord observed that $F\colon [0, \infty) \to [0, \infty)$ defined by

$$F(x) = \begin{cases} 0 & \text{if } 0 \leq x \leq t - 1 \\ \binom{x}{t} = x(x-1)\cdots(x-t+1)/t! & \text{if } x > t - 1 \end{cases}$$

is a convex function. Hence $\frac{1}{n}\sum_{k=1}^{n} F(v_k) \geq F\left(\frac{1}{n}\sum \#B_i\right) = F(\lambda)$. Thus when $1 \leq t \leq \lambda$ the lower bound can be improved by replacing $(\lambda - t + 1)$ with $\binom{\lambda}{t}$. In a more detailed solution, John H. Lindsey II obtained both latter extensions.

Solved also by R. Barbara (Lebanon), R. J. Chapman (U. K.), V. Hernández (Spain), J. H. Lindsey II, O. P. Lossers (The Netherlands), R. J. Simpson (Australia), P. Tracy, A. N. 't Woord (The Netherlands), USA Problem Group, and the proposer.

A Lower Bound on Correlation

10384 [1994, 474]. *Proposed by Franklin Kemp, East Texas State University, Commerce, TX.* Suppose $x_1 < x_2 < \cdots < x_n$ and $y_1 < y_2 < \cdots < y_n$. Define the correlation coefficient r in the usual way:

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \cdot \sum_i (y_i - \overline{y})^2}}$$

where $\overline{x}$ and $\overline{y}$ are the average values of the x_i and y_i, respectively, and the sums run from 1 to n. Show that $r \geq 1/(n-1)$.

Solution by Robin J. Chapman, University of Exeter, Exeter, U. K.. By replacing x_i by $x_i - \overline{x}$ and y_i by $y_i - \overline{y}$ we may assume that $\overline{x} = \overline{y} = 0$. Let

$$C = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \colon x_1 < x_2 < \cdots < x_n,\ x_1 + x_2 + \cdots + x_n = 0\}.$$

Denote the Euclidean inner product of two elements $\mathbf{u}, \mathbf{v}$ of $\mathbb{R}^n$ by $\mathbf{u} \cdot \mathbf{v}$. Let $\mathbf{u}_j = (1, 1, \ldots, 1, 0, \ldots, 0)$ (with j 1's) for $0 \leq j \leq n$ and $\mathbf{v}_j = (1/n)\mathbf{u}_n - (1/j)\mathbf{u}_j$ for $1 \leq j \leq n-1$. Note that each $\mathbf{v}_j$ is in C.

Given $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in C$, a computation shows that

$$\mathbf{x} = \sum_{j=1}^{n-1} j\left(x_{j+1} - x_j\right)\mathbf{v}_j,$$

and so $\mathbf{x} = \sum_i s_i \mathbf{v}_i$ with each $s_i > 0$. Similarly, given $\mathbf{y} \in C$, $\mathbf{y} = \sum_i t_i \mathbf{v}_i$ with each $t_i > 0$.

Now let $a_{ij} = \mathbf{v}_i \cdot \mathbf{v}_j$. If $i \leq j$ then

$$a_{ij} = \left(\frac{1}{n}\mathbf{u}_n - \frac{1}{i}\mathbf{u}_i\right) \cdot \left(\frac{1}{n}\mathbf{u}_n - \frac{1}{j}\mathbf{u}_j\right) = \frac{1}{j} - \frac{1}{n},$$

independent of i, and so

$$\frac{a_{ij}}{\sqrt{a_{ii}a_{jj}}} = \frac{a_{jj}}{\sqrt{a_{ii}a_{jj}}} = \sqrt{\frac{a_{jj}}{a_{ii}}} = \sqrt{\frac{j^{-1} - n^{-1}}{i^{-1} - n^{-1}}} \geq \sqrt{\frac{(n-1)^{-1} - n^{-1}}{1 - n^{-1}}} = \frac{1}{n-1}.$$

Thus, the desired inequality is valid among the $\mathbf{v}_i$. Using this, we find

$$\mathbf{x} \cdot \mathbf{y} = \sum_i \sum_j a_{ij} s_i t_j \geq \frac{1}{n-1} \sum_i \sum_j s_i t_j \sqrt{a_{ii}a_{jj}} = \frac{1}{n-1} \sum_i s_i \sqrt{a_{ii}} \sum_j t_j \sqrt{a_{jj}}.$$

Since the Cauchy-Schwarz inequality gives $a_{ij} = \mathbf{v}_i \cdot \mathbf{v}_j \leq \sqrt{a_{ii}a_{jj}}$, we have

$$\mathbf{x} \cdot \mathbf{x} = \sum_i \sum_j s_i s_j a_{ij} \leq \sum_i \sum_j s_i s_j \sqrt{a_{ii}a_{jj}} = \left(\sum_i s_i \sqrt{a_{ii}}\right)^2,$$

and similarly

$$\mathbf{y} \cdot \mathbf{y} \leq \left(\sum_j t_j \sqrt{a_{jj}}\right)^2.$$

Thus, we conclude that

$$r = \frac{\mathbf{x} \cdot \mathbf{y}}{\sqrt{\mathbf{x} \cdot \mathbf{x}}\sqrt{\mathbf{y} \cdot \mathbf{y}}} \geq \frac{1}{n-1}$$

as required.

This argument can be refined to show that the inequality is strict for $n \geq 3$, but $1/(n-1)$ is the best possible constant.

Editorial comment. As some solvers remarked, it suffices to require only $x_1 \le x_2 \le \cdots \le x_n$ with $x_1 < x_n$ (and similarly for the y_j), but one must demand that $n > 1$.

Can Anh Minh noted that the result appears in Christopher Bradley and Nick Lord, Computing Spearman's rank and the product-moment correlation coefficients, *The Mathematical Gazette* 77 (1993), 84–88.

Solved also by P. G. Kirmser, O. Krafft & M. Schaefer (Germany), R. Kreminski, J. H. Lindsey II, L. E. Mattics, D. P. Walsh, T. White, and the proposer.

An Application of Snake Oil

10388 [1994, 474]. *Proposed by E. Sparre Andersen and Mogens Esrom Larsen, Københavns Universitet, København, Denmark.* Find

$$\sum_{k=0}^{n}\binom{n}{k}\binom{\frac{n-3}{4}-\frac{k}{2}+p}{2p}$$

where n and p are positive integers.

Solution by H. van Haeringen, Delft University of Technology, Delft, The Netherlands. The proof uses methods described in H. S. Wilf, *generatingfunctionology* (second ed.), Academic Press, 1994, which we denote by [g]. Let $S_n(p)$ be the sum in question; we prove that

$$S_n(p) = (-1)^p 2^{n-4p}\binom{2p-\frac{n+1}{2}}{p}.$$

First we use the "Snake Oil method" [g, p. 118–130] to prove that

$$\sum_{p=0}^{\infty}(-4\sin^2\theta)^p\binom{a+p}{2p} = \frac{\cos(1+2a)\theta}{\cos\theta}. \tag{1}$$

Letting L_a be the left-hand side of (1), we have

$$\begin{aligned}\sum_{a=0}^{\infty}L_a z^a &= \sum_{p=0}^{\infty}(-4\sin^2\theta)^p\sum_{a=0}^{\infty}\binom{a+p}{2p}z^a = \sum_{p=0}^{\infty}(-4\sin^2\theta)^p\frac{z^p}{(1-z)^{2p+1}}\\ &= \frac{1-z}{(1-z)^2+4z\sin^2\theta} = \frac{1-z}{1-2z\cos 2\theta+z^2}\\ &= \frac{1}{2\cos\theta}\left(\frac{e^{i\theta}}{1-ze^{2i\theta}}+\frac{e^{-i\theta}}{1-ze^{-2i\theta}}\right) = \sum_{a=0}^{\infty}\frac{\cos(1+2a)\theta}{\cos\theta}z^a.\end{aligned}$$

This proves (1) when a is nonnegative integer. On both sides, the coefficient of θ^n in the power series expansion is a polynomial in a; therefore (1) holds for all real numbers a.

By (1),

$$\sum_{p=0}^{\infty}(-4\sin^2\theta)^p S_n(p) = \sum_{k=0}^{n}\binom{n}{k}\frac{\cos\big(1+(n-3)/2-k\big)\theta}{\cos\theta}.$$

This sum equals $2^{(n-1)/2}\big(1+\cos\theta\big)^{(n+1)/2}/\cos\theta$ by a straightforward computation using $\cos u = \big(e^{iu}+e^{-iu}\big)/2$ and the binomial theorem. Letting $u=\sin^2\theta$, we obtain

$$\begin{aligned}\sum_{p=0}^{\infty}(-4u)^p S_n(p) &= 2^{(n-1)/2}\frac{\big(1+\sqrt{1-u}\big)^{(n+1)/2}}{\sqrt{1-u}}\\ &= \frac{2^n}{\sqrt{1-u}}\left(\frac{1-\sqrt{1-u}}{u/2}\right)^{-(n+1)/2} = 2^n\sum_{p=0}^{\infty}\left(\frac{u}{4}\right)^p\binom{2p-\frac{n+1}{2}}{p},\end{aligned}$$

where the last step follows by a well-known identity [**g**, equation 2.5.15]. The formula for $S_n(p)$ follows by equating coefficients of u^p on each side.

Editorial comment. The "Snake Oil method" to evaluate a sum involves forming a generating function over a fixed parameter a, interchanging the order of summation, and then performing the inner sum. The proposers' proof uses formulas that they developed to study more general sums; van Haeringen also gave two proofs using hypergeometric series and a fourth proof based on a recurrence for $S_n(p)$. Herbert S. Wilf and István Nemes provided computer proofs based on the algorithm of D. Zeilberger, A fast algorithm for proving terminating hypergeometric identities, *Discrete Math.* 80 (1990), 207–211. Wilf notes that the computer method now provides a general approach to such sums.

Solved also by I. Nemes (Austria), H. S. Wilf and the proposers.

Tower of Power

10389 [1994, 574]. *Proposed by Raphael M. Robinson, University of California, Berkeley, CA.* Find all solutions of the equation

$$a_1^{a_2^{\cdot^{\cdot^{\cdot^{a_m}}}}} = b_1^{b_2^{\cdot^{\cdot^{\cdot^{b_n}}}}}$$

where $m \geq 1$, $n \geq 1$, the a_i and b_i are integers with $2 \leq a_i \leq 4$ and $2 \leq b_i \leq 4$, and $a_1 \neq b_1$.

Composite solution by all solvers, with a generalization by the editors. We show more generally that for $N \geq 2$ there are finitely many solutions with $m \geq 1$, $n \geq 1$, $a_1 \neq b_1$, and each $a_i, b_i \in \{2, \ldots, N\}$. In fact, the upper bound N is needed only for $1 \leq i \leq 3$, and the solutions for each N can be computed effectively. We represent the desired equation using the notation

$$[\![a_1, \ldots, a_m]\!] = [\![b_1, \ldots, b_n]\!]. \tag{1}$$

By unique factorization, it is clear that a_1 and b_1 have the same prime factors. We consider each such a_1 and b_1 with $a_1 \neq b_1$ and $2 \leq a_1, b_1 \leq N$. Since $a_1 \neq b_1$, there is a prime p such that $p^r \| a_1$ and $p^s \| b_1$ where r and s are distinct positive integers (the notation $p^r \| a$ means that p^r is the largest power of p dividing a). We may assume that $r < s$. By considering the largest power of p that divides both sides of (1), we deduce that

$$r[\![a_2, \ldots, a_m]\!] = s[\![b_2, \ldots, b_n]\!]. \tag{2}$$

The cases in (2) where a_2 and b_2 have different sets of prime factors can be solved easily. If q is a prime dividing b_2 but not a_2, then $[\![q, b_3, \ldots, b_n]\!]$ divides r. Thus $[\![b_3, \ldots, b_n]\!] \leq \log_q r$, and $[\![b_1, \ldots, b_n]\!] \leq [\![N, N, \log_q r]\!]$. In particular, there are finitely many such values.

We may suppose then that a_2 and b_2 have the same prime factors. Since $r < s$, there are nonnegative integers r' and s' with $r' < s'$ and a prime q such that $q^{r'} \| r$ and $q^{s'} \| s$. From (2), we must have $q | a_2$, and hence also $q | b_2$, since a_2 and b_2 have the same prime factors. Let c and d be positive integers such that $q^c \| a_2$ and $q^d \| b_2$. By considering the largest power of q that divides both sides of (2), we deduce that

$$r' + c[\![a_3, \ldots, a_m]\!] = s' + d[\![b_3, \ldots, b_n]\!],$$

and thus $c[\![a_3, \ldots, a_m]\!] > d[\![b_3, \ldots, b_n]\!]$.

Since $2 \leq a_3, b_3 \leq N$, we need only find all nonnegative integer solutions x and y to the equations $cu^x - dv^y = w$ where c, d, u, v, and w are integers satisfying

$$1 \leq c, d \leq \log_2 N, \quad 1 \leq w \leq \log_2 \log_2 N, \quad \text{and} \quad 2 \leq u, v \leq N.$$

That this can be done follows, for example, from Theorem 1.1 of Shorey and Tijdeman, *Exponential Diophantine Equations*, Cambridge Tracts in Math. 87, Cambridge, 1986.

When $N = 4$, the problem reduces to solving the diophantine equations $2^x - 3^y = 1$ (with solution set $\{(1, 0), (2, 1)\}$) and $3^x - 2^y = 1$ (with solution set $\{(1, 1), (2, 3)\}$). These solutions are well known. They yield eleven solutions to the proposed problem:

$$2^2 = 4, \quad 2^4 = 4^2, \quad 2^{2^2} = 4^2, \quad 2^{2^{2^2}} = 4^{2^3}, \quad 2^{2^3} = 4^{2^2}, \quad 2^{2^3} = 4^4,$$
$$2^{2^{3^2}} = 4^{2^{2^3}}, \quad 2^{2^{3^2}} = 4^{4^{2^2}}, \quad 2^{2^{3^2}} = 4^{4^4}, \quad 2^{2^4} = 4^{2^3}, \quad \text{and} \quad 2^{4^2} = 4^{2^3}.$$

Editorial comment. All solvers used the same approach, restricted to the special case $N = 4$, but four solvers omitted one or more of the solutions.

Solved by R. Barbara (Lebanon), L. Cagliero (Argentina), R. J. Chapman (U. K.), J. Christopher, Z. Franco, H. W. Guggenheimer, R. Holzsager, N. Komanda, J. Kuliyev, J. H. Lindsey II, O. P. Lossers (The Netherlands), J. Merickel, C. A. Minh, A. Pedersen (Denmark), J. Rostand (Canada), T. Tran, T. White, A. N. 't Woord (The Netherlands), NSA Problems Group, Prague Problem Solution Group (Czech Republic), and the proposer.

Growing Inequalities

10391 [1994, 574]. *Proposed by Emre Alkan (student), Bosphorus University, İstanbul, Turkey, and the editors.* If $a_1, a_2, \ldots, a_n$ are real numbers with $a_1 \geq a_2 \geq \cdots \geq a_n$, and if ϕ is a convex function defined on the closed interval $[a_n, a_1]$, then

$$\sum_{k=1}^{n} \phi(a_k)a_{k+1} \geq \sum_{k=1}^{n} \phi(a_{k+1})a_k$$

with the convention that $a_{n+1} = a_1$.

Solution I by Zachary Franco, Butler University, Indianapolis, IN. For $n = 1$ and 2 the inequality is trivial. Let $a_1, a_2, \ldots$ be an infinite nonincreasing sequence of real numbers and let $S_n = (\phi(a_n)a_1 - \phi(a_1)a_n) + \sum_{k=1}^{n-1}(\phi(a_k)a_{k+1} - \phi(a_{k+1})a_k)$. We must show $S_n \geq 0$.

By Jensen's inequality for convex functions, $\phi(\alpha x + \beta y) \leq \alpha\phi(x) + \beta\phi(y)$. Put

$$\alpha = \frac{a_n - a_{n+1}}{a_1 - a_{n+1}}, \quad \beta = \frac{a_1 - a_n}{a_1 - a_{n+1}}, \quad x = a_1, \quad \text{and} \quad y = a_{n+1}$$

and simplify to obtain

$$(a_n - a_{n+1})\phi(a_1) + (a_1 - a_n)\phi(a_{n+1}) + (a_{n+1} - a_1)\phi(a_n) \geq 0.$$

The left side of this inequality is precisely $S_{n+1} - S_n$, which implies that S_n is a nondecreasing sequence and hence must always be nonnegative.

Solution II by Richard Holzsager, American University, Washington, DC. Fix n. The two sides are clearly equal for $\phi(x) = 1$ or $\phi(x) = x$, and since both sides are linear in ϕ, we can subtract off the linear function that agrees with ϕ at the endpoints and reduce to the case where $\phi(a_1) = \phi(a_n) = 0$. Eliminating the terms with these factors allows the desired inequality to be rewritten $\sum_{k=2}^{n-1} \phi(a_k)\big(a_{k+1} - a_{k-1}\big) \geq 0$. But since ϕ is convex and zero at the endpoints, each $\phi(a_k)$ is nonpositive, as is the other factor. The inequality follows.

Editorial comment. The left side of the three term inequality used in Solution I may be interpreted as the value of the well-known determinant formula for the area of the triangle with vertices $\big(a_i, \phi(a_i)\big)$, $i = 1, n, n+1$. The convexity condition determines the orientation of the triangle and hence the sign of this determinant. The solution of O. P. Lossers used this interpretation.

The National Security Agency Problems Group gave a more visual form of Solution II. They began their solution by subtracting $\sum_{k=1}^{n} \phi(a_k)a_k$ from each side of the desired

inequality. The resulting sums were then written with factors of $\delta_k = a_k - a_{k+1}$ in each term. This led to sums arising in the trapezoidal approximation to $\int_{a_n}^{a_1} \phi(x)\,dx$.

An algebraic form of this problem appeared in *Crux Mathematicorum* as Problem 10-3, and it was discussed in detail in the "Olympiad Corner" section of that journal [1980, 106–108 and 129–130].

Solved also by J. Alvarez (Spain), D. Borwein (Canada), R. J. Chapman (U. K.), J. H. Chyung, R. Ehrenborg (Canada), E. Escalona-Fernandez, J.-P. Grivaux (France), H. van Haeringen (The Netherlands), E. A. Herman, K. Hinderer & M. Stieglitz (Germany), J. Howard, F.-A. Izadi (Iran), G. Keselman, B. S. Kim (Korea), P. G. Kirmser, J. H. Lindsey II, O. P. Lossers (The Netherlands), P. McCartney, K. D. McLenithan, J. Merickel, C. A. Minh, A. Pedersen (Denmark), K. Schilling, M. Vowe (Switzerland), A. N. 't Woord (The Netherlands), NSA Problems Group, Prague Problem Solution Group (Czech Republic), USA Problem Group, and the WMC Problems group.

The Effect of an Alternating Series

10398 [1994, 682]. *Proposed by Leroy Quet, Denver, CO.* Show that

$$\sum_{n=1}^{\infty}\sum_{m=1}^{n} \frac{1}{m\cdot(n+1)!} = \sum_{n=1}^{\infty}\sum_{m=1}^{n} \frac{(-1)^{n+1}e}{m\cdot(n+1)!}.$$

Solution by Michael Vowe, Therwil, Switzerland. Let H_n be the sum of the first n terms of the harmonic series, so $H_n = \sum_{m=1}^{n} 1/m = \int_0^1 (t^n - 1)/(t-1)\,dt$. Consider the function f represented by the absolutely convergent infinite series

$$f(x) = \sum_{n=1}^{\infty} \frac{x^n}{(n+1)!}H_n = \int_0^1 \frac{1}{t-1}\sum_{n=1}^{\infty}\frac{t^n x^n - x^n}{(n+1)!}\,dt = \int_0^1 \frac{e^{tx} - 1 - te^x + t}{t(t-1)x}\,dt.$$

We have $f(1) = \sum_{n=1}^{\infty} H_n/(n+1)! = \int_0^1 (e^t - 1 - te + t)/(t(t-1))\,dt$.

Consider also the function g represented by the absolutely convergent infinite series

$$g(x) = \sum_{n=1}^{\infty} \frac{(-x)^{n+1}e^x}{(n+1)!}H_n = \int_0^1 \frac{e^x}{t-1}\sum_{n=0}^{\infty}\frac{(-xt)^{n+1}t^{-1} - (-x)^{n+1}}{(n+1)!}\,dt$$
$$= \int_0^1 \frac{e^x}{t-1}\left(\frac{e^{-xt}-1}{t} - (e^{-x}-1)\right)dt = \int_0^1 \frac{e^{(1-t)x} + (t-1)e^x - t}{(t-1)t}\,dt.$$

Setting $s = 1-t$ yields

$$g(1) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}e}{(n+1)!}H_n = \int_0^1 \frac{(e^s - se + s - 1)}{s(s-1)}\,ds = f(1),$$

as desired.

Solved also by J. Anglesio (France), D. Beckwith, D. Borwein (Canada), P. Bracken (Canada), D. Bradley, E. Braune (Austria), P. Budney, R. J. Chapman (U. K.), H. Chen, D. A. Darling, R. Fraties, C. Georghiou (Greece), M. .L. Glasser, E. Hertz, M. Hoffman, R. Holzsager, G. L. Isaacs, G. Keselman, B. G. Klein, J. H. Lindsey II, O. P. Lossers (The Netherlands) A. R. Miller, D. K. Nester, A. Nijenhuis, A. E. C. Núñez (Colombia), H. Prodinger (Austria), R. Sprugnoli, R. Stong, D. B. Tyler, C. Y. Yildirim (Turkey), NSA Problems Group, USA Problems Group, WMC Problems Group, and the proposer.

Bounding a Binomial Sum

10400 [1994, 682]. *Proposed by Itshak Borosh, Douglas Hensley, and Arthur M. Hobbs, Texas A&M University, College Station, TX, and Anthony Evans, Wright State University, Dayton, OH.* Determine the set of all pairs (n, t) of integers with $0 \le t \le n$ and

$$\sum_{k=0}^{t}\binom{n}{k} < \frac{n^t}{t!}.$$

Solution by Stephen M. Gagola, Jr., Kent State University, Kent, OH. The inequality holds precisely when $t = 4$ and $n \geq 7$, when $t = 5$ and $n \geq 8$, and when $6 \leq t \leq n$.

When $0 \leq t \leq 3$, the left side of the inequality is a polynomial in n of degree t whose leading term is $n^t/t!$ and whose remaining nonzero terms have positive coefficients. Thus the inequality cannot hold.

When $t = 4$, the inequality (times 24) is $n^4 - 2n^3 + 11n^2 + 14n + 24 < n^4$. This fails for $n \in \{4, 5, 6\}$. For $n \geq 7$, we have $11n^2 + 14n + 24 < 11n^2 + 2n^2 + n^2 = 14n^2 \leq 2n^3$, and the inequality holds.

When $t = 5$, the inequality (times 120) is $n^5 - 5n^4 + 25n^3 + 5n^2 + 94n + 120 < n^5$. This fails for $n = 5$. For $n \geq 6$, we have $25n^3 + 5n^2 + 94n + 120 < 25n^3 + n^3 + 3n^3 + n^3 = 30n^3 \leq 5n^4$, and the inequality holds.

When $n = t \geq 6$, the inequality becomes $2^n n! < n^n$, which holds inductively for $n \geq 6$. With these statements as basis, we now complete the proof by induction on n. For the induction step, we consider $n > t$ and use the induction hypothesis to compute

$$\sum_{k=0}^{t}\binom{n}{k} = \sum_{k=0}^{t}\binom{n-1}{k} + \sum_{k=1}^{t}\binom{n-1}{k-1} < \frac{(n-1)^t}{t!} + \frac{(n-1)^{t-1}}{(t-1)!}$$
$$= \frac{(n-1)^t}{t!}\left(1 + \frac{t}{n-1}\right) < \frac{(n-1)^t}{t!}\left(1 + \frac{1}{n-1}\right)^t = \frac{n^t}{t!}.$$

Solved also by J. Anglesio (France), M. Ascher, J. T. Bruening, J. Christopher, J. S. Frame, J. Graham & J. B. Klerlein & S. Sportsman, R. Holzsager, G. L. Isaacs, N. Komanda, J. Lauret (Argentina), J. H. Lindsey II, G. T. Lee (Canada), O. P. Lossers (Netherlands), G. Myerson (Australia), I. Nemes (Austria), V. Novakov (Bulgaria), R. Stong, D. B. Tyler, NSA Problems Group, WMC Problems Group, and the proposer.

The Average Order of an Element in a Group

10410 [1994, 911]. *Proposed by Frank Schmidt, Arlington, VA.* Let G be a finite group. Define $a(G)$ to be the average order of an element of G. If $|G| \neq 1$, can $a(G)$ be an integer?

Solution by Peter L. Montgomery, San Rafael, CA. Yes. First, we make two observations.

(1) If G_1 and G_2 are finite groups and $\gcd(|G_1|, |G_2|) = 1$, then $a(G_1 \times G_2) = a(G_1)a(G_2)$.

(2) If p is a prime and $C(p)$ denotes the cyclic group of order p, then $a(C(p)) = (p^2 - p + 1)/p$.

To prove (1), observe that the order of an element (g_1, g_2) of $G_1 \times G_2$ is always the least common multiple of the orders of g_1 and g_2. The assumption guarantees that this is the product of the orders of g_1 and g_2. Thus, (1) follows by grouping the terms in the sum of the orders. To prove (2), note that $C(p)$ has $p - 1$ elements of order p and 1 element of order 1.

We look for a cyclic group $G = C(p_1) \times C(p_2) \times \cdots \times C(p_n)$, where the p_i are distinct primes. This gives the example we seek if $p_1 p_2 \cdots p_n$ divides $(p_1^2 - p_1 + 1)(p_2^2 - p_2 + 1) \cdots (p_n^2 - p_n + 1)$. Any prime factor of the latter product (other than 3) must be congruent to 1 modulo 3, so try $p_1 = 7$. Continuing: $p_1^2 - p_1 + 1 = 43$, so try $p_2 = 43$; $p_2^2 - p_2 + 1 = 1807 = 13 \times 139$, so try $p_3 = 13$; $p_3^2 - p_3 + 1 = 157$, so try $p_4 = 157$; $p_4^2 - p_4 + 1 = 24493 = 7 \times 3499$. Since we now have a multiple of p_1, the cycle is complete. With $G = C(7) \times C(13) \times C(43) \times C(157) = C(614341)$, $a(G)$ is the integer $139 \times 3499 = 486361$.

A noncyclic abelian example can be found in a similar way. Let $G = C(13) \times C(13) \times C(23)$ (of order 3887). Then $a(G) = (13^3 - 13 + 1)(23^2 - 23 + 1)/3887 = 285$.

Editorial comment. Douglas B. Tyler noted that there are at least 100 examples of such groups, and raised the question of whether there are infinitely many examples. One example can often spawn others, since (1) implies that if $a(G)$ is an integer, d is any divisor of $a(G)$ relatively prime to $|G|$, and H is any group of order d, then $a(G \times H)$ is an integer.

Nilpotent groups are characterized as direct products of their Sylow subgroups (see Marshall Hall, Jr., *The Theory of Groups*, Macmillan, 1959, ch. 10). Knowing $a(G)$ for all groups G of prime-power order allows a systematic study of this problem for nilpotent groups. In particular, one can verify that the smallest nilpotent group G for which $a(G)$ is an integer is the abelian group of order 3887 that has already been mentioned.

The examples in the selected solution were most frequently cited. Another popular example was constructed by taking the direct product of $C(5)$ with the nonabelian group of order 21. John H. Lindsey II proved that this is the smallest group with the desired property. The proof consists of showing that any smaller example must be nilpotent by using the fact that smaller numbers have the form $p^a q$, and two more simple observations:

(3) The sum of the orders of the elements of a group is always odd.

(4) If p divides $|G|$, then the sum of the orders of the elements of G is congruent modulo p to the sum of the orders of elements in the centralizer of a p-Sylow subgroup.

First, let a p-Sylow subgroup act on G by conjugation. Elements in the same orbit have the same order, so only the fixed points contribute to the sum modulo p. This gives (4). Indeed, this centralizer further splits as a direct product of a group whose order is a power of p and one whose order is relatively prime to p. One could further restrict to the latter factor, since the other elements have orders that are multiples of p. The proof of (3) is similar, starting from the fact that an element has the same order as its inverse.

Solved also by R. Holzsager, J. Lauret (Argentina), J. H. Lindsey II, J. H. Nieto (Venezuela), R. Stong, D. B. Tyler, A. N. 't Woord (The Netherlands), and the proposer.

A Bernoulli Convolution

10416 [1994, 912]. *Proposed by Kwang-Wu Chen (student), National Chung Cheng University, Chia-Yi, Taiwan, Republic of China.* The Bernoulli numbers B_n (for $n = 0, 1, 2, \ldots$) are defined by

$$\frac{t}{e^t - 1} = \sum_{n=0}^{\infty} \frac{B_n t^n}{n!},$$

which converges for $|t| < 2\pi$. Also, for each nonnegative integer n, the Bernoulli polynomial $B_n(x)$ is defined by

$$B_n(x) = \sum_{k=0}^{n} \binom{n}{k} B_{n-k} x^k.$$

For integer $m \geq 1$ and arbitrary constants α and β, prove

$$\sum_{k=0}^{m} \binom{m}{k} B_k(\alpha) B_{m-k}(\beta) = -(m-1) B_m(\alpha+\beta) + m(\alpha+\beta-1) B_{m-1}(\alpha+\beta).$$

Solution by David Borwein, University of Western Ontario, London, Ontario, Canada. Let L_m and R_m denote respectively the left side and the right side of the desired equation. It is well known and easily verified that

$$\sum_{n=0}^{\infty} \frac{B_n(x) t^n}{n!} = \frac{t e^{xt}}{e^t - 1} \text{ for } |t| < 2\pi.$$

For $|t| < 2\pi$, we thus have

$$\sum_{m=0}^{\infty} \frac{L_m t^m}{m!} = \left(\sum_{k=0}^{\infty} \frac{B_k(\alpha) t^k}{k!} \right) \left(\sum_{l=0}^{\infty} \frac{B_l(\beta) t^l}{l!} \right) = \frac{t e^{\alpha t}}{e^t - 1} \cdot \frac{t e^{\beta t}}{e^t - 1} = \frac{t^2 e^{(\alpha+\beta)t}}{(e^t - 1)^2},$$

and

$$\sum_{m=0}^{\infty}\frac{R_m t^m}{m!} = -\sum_{m=0}^{\infty}\frac{(m-1)B_m(\alpha+\beta)t^m}{m!} + (\alpha+\beta-1)\sum_{m=1}^{\infty}\frac{B_{m-1}(\alpha+\beta)t^m}{(m-1)!}$$
$$= -t^2\frac{d}{dt}\left(\frac{e^{(\alpha+\beta)t}}{e^t-1}\right) + (\alpha+\beta-1)\frac{t^2e^{(\alpha+\beta)t}}{e^t-1}$$
$$= -(\alpha+\beta)\frac{t^2e^{(\alpha+\beta)t}}{e^t-1} + \frac{t^2e^{(\alpha+\beta)t}e^t}{(e^t-1)^2} + (\alpha+\beta-1)\frac{t^2e^{(\alpha+\beta)t}}{e^t-1}$$
$$= \frac{t^2e^{(\alpha+\beta)t}e^t}{e^t-1} - \frac{t^2e^{(\alpha+\beta)t}}{e^t-1} = \frac{t^2e^{(\alpha+\beta)t}}{(e^t-1)^2}.$$

It follows that $L_m = R_m$ for $m \geq 1$.

Editorial comment. Carl Axness pointed out that the identity appeared as series (50.11.2) in Eldon R. Hansen, *A Table of Series and Products*, Prentice Hall, 1975. Bruce Dearden and F. T. Howard independently generalized the identity to Bernoulli polynomials of higher order. With $B_n^{(m)}(x)$ defined by $\sum_{n=0}^{\infty} B_n^{(m)}(x)t^n/n! = t^m e^{xt}/(e^t-1)^m$, they proved that

$$\sum_{k=0}^{m}\binom{m}{k}B_k^{(p)}(\alpha)B_{m-k}^{(q)}(\beta)$$
$$= -\left(\frac{m}{p+q-1}-1\right)B_m^{(p+q-1)}(\alpha+\beta) + m\left(\frac{\alpha+\beta}{p+q-1}-1\right)B_{m-1}^{(p+q-1)}(\alpha+\beta),$$

which reduces to the present identity when $p = q = 1$.

Solved also by E. S. Andersen & M. E. Larsen (Denmark), J. Anglesio (France), T. M. Apostol, C. Axness, D. Bradley, J. C. Carey, R. J. Chapman (U.K.), B. Dearden, R. Ehrenborg (Canada), C. Georghiou (Greece), M. Hauss (Germany), F. T. Howard, J. H. Lindsey II, O. P. Lossers (Netherlands), I. Nemes (Austria), A. Pedersen (Denmark), R. Pirastu (Austria), A. Stenger, A. N. 't Woord (Netherlands), M. Vowe (Switzerland), D. Zeilberger, Anchorage Problem Solutions Group, NSA Problems Group, WMC Problems Group, and the proposer.

A Consequence of Fermat's Little Theorem

10417 [1994, 1013]. *Proposed by Proposed by Charles Vanden Eynden, Illinois State University, Normal, IL.* Characterize the positive integers m such that

$$m^n \equiv 1 \pmod{n} \implies m \equiv 1 \pmod{n}.$$

Solution by Todd H. Trimble, Loyola University, Chicago, IL. The only such integers are $m = 1, 2$. If $m > 2$, then the binomial expansion yields

$$m^{(m-1)^2} \equiv \big(1+(m-1)\big)^{(m-1)^2} \equiv 1 \pmod{(m-1)^2},$$

but $m \not\equiv 1 \pmod{(m-1)^2}$.

Since $m = 1$ works trivially, it suffices to show that the hypothesis never holds when $m = 2$. Suppose $2^n \equiv 1 \pmod{n}$ for some $n > 1$. Then $2^n \equiv 1 \pmod{p}$, where p is the least prime dividing n. Clearly $p \neq 2$, so p is an odd prime. Since n is not divisible by any prime smaller than p, it is relatively prime to $p-1$. By the Euclidean algorithm, there exist integers a, b such that $an + b(p-1) = 1$. Using Fermat's Little Theorem, we now obtain the contradiction $2^1 = 2^{na}2^{(p-1)b} \equiv 1 \pmod{p}$.

Editorial comment. Gerry Myerson observed that the case $m = 2$ is Problem A-5 on the 1972 Putnam exam (this MONTHLY 80 (1973), 1017–1028) and showed that the implication fails also for negative m.

Solved also by R. Barbara (Lebanon), J. T. Bruening, D. Callan, R. J. Chapman (U. K.), J. Christopher, G. Ehrlich, R. Holzsager, M. Hudelson , J. H. Lindsey II, O. P. Lossers (The Netherlands), G. Myerson (Australia), A. Pedersen (Denmark), M. Reid, A. A. Tarabay (Lebanon), D. B. Tyler, H. Widmer (Switzerland), A. N. 't Woord (The Netherlands), NSA Problems Group, and the proposer.

A Sequence of Polynomials with Positive Coefficients

10420 [1994, 1014]. *Proposed by C. R. Selvaraj and S. Selvaraj, Pennsylvania State University, Shenango Campus, Sharon, PA.* Let

$$g_i(n) = \sum_{k=i}^{\infty} \frac{k-i+1}{k!}\left((n+2)^k - 2e(n+1)^k + e^2 n^k\right).$$

Prove that, for all $i > 1$, $g_i(n)$ is a polynomial in n of degree $i-2$ and $g_i(n) \geq 0$ for all $n \in \mathbb{N}$ and $i \in \mathbb{N}$.

Composite solution by many solvers. We prove something stronger: For $i \geq 2$, g_i is a polynomial of degree $i-2$ with all coefficients positive. Since $\sum_{k=0}^{\infty}(k-i+1)x^k/k! = (x-i+1)e^x$, extending the sum given for $g_i(n)$ over all nonnegative k yields

$$(n-i+3)e^{n+2} - 2e(n-i+2)e^{n+1} + e^2(n-i+1)e^n,$$

which is 0 for all n. The term with $k = i-1$ is also 0, so

$$g_i(n) = \sum_{k=0}^{i-2} \frac{i-1-k}{k!}\left((n+2)^k - 2e(n+1)^k + e^2 n^k\right).$$

Thus g_i is a polynomial of degree $i-2$ with leading coefficient $(1-2e+e^2)/(i-2)! = (e-1)^2/(i-2)!$.

We complete the proof using induction. We have shown that $g_2(x)$ is the constant $(e-1)^2$. Suppose $i > 2$ and the coefficients of g_{i-1} are positive. The constant term of g_i is $g_i(0) = \sum_{k=i}^{\infty}(k-i+1)(2^k-2e)/k!$, which is a sum of positive numbers. Differentiating the polynomial g_i termwise yields $g_i' = g_{i-1}$. By the induction hypothesis, we conclude that the rest of the coefficients are also positive.

Editorial comment. Several solvers established the generating function identity

$$\sum_{i=2}^{\infty} g_i(n)t^i = t^2\left(\sum_{j=1}^{\infty}\frac{1+t+t^2+\cdots+t^{j-1}}{j!}\right)^2 e^{nt},$$

from which it follows readily that g_i is a polynomial of degree $i-2$ with positive coefficients.

Solved by J. Anglesio (France), D. W. Bailey, D. A. Beckwith, D. M. Bradley, R. J. Chapman (England), C. Georghiou (Greece), J. H. Lindsey II, O. P. Lossers (The Netherlands), D. K. Nester, I. A. Sakmar (Turkey), J. H. Steelman, A. Stenger, D. B. Tyler, the NSA Problems Group, the WMC Problems Group, and the proposers.

A Surprisingly Simple Summation Solution

10424 [1995, 70]. *Proposed by Ira Gessel, Brandeis University, Waltham, MA.* Evaluate the sum

$$\sum_{0\leq k\leq n/3} 2^k \frac{n}{n-k}\binom{n-k}{2k}.$$

Solution I by E. Sparre Andersen and Mogens Esrom Larsen, University of Copenhagen, Copenhagen, Denmark. The desired sum $s(n)$ is $2^{n-1}+\cos(n\pi/2)$. This follows from

$$s(n) - 2s(n-1) + s(n-2) - 2s(n-3) = 0, \tag{1}$$

which has the general solution $s(n) = A2^n + Bi^n + C(-i)^n$. Since $s(1) = s(2) = 1$ and $s(3) = 4$, we obtain $A = B = C = 1/2$.

To establish (1), we express the sum as $s(n) = u(n) + v(n)$, where

$$u(n) = \sum_{0 \le k \le n/3} 2^k \binom{n-k}{2k}, \qquad v(n) = \sum_{1 \le k \le n/3} 2^{k-1} \binom{n-k-1}{2k-1}.$$

Applying the binomial recurrence yields the recurrences

$$u(n) = u(n-1) + 2v(n) \qquad \text{and} \qquad v(n) = v(n-1) + u(n-3).$$

Repeated application of these relations shows that u and v satisfy (1), and thus their sum also satisfies (1).

Solution II by Thomas Honold, Technische Universität München, München, Germany. The answer is 2^{n-1} when n is odd and $2^{n-1} + (-1)^{n/2}$ when n is even. Using the notation of Solution I, we form the generating function $f(x) = 1 + \sum_{n \ge 1} s(n)x^n$. Letting $n = 3k + m$ and using $(1-x)^{-s} = \sum_{m \ge 0} \binom{m+s-1}{s-1} x^m$, we obtain

$$1 + \sum_{n \ge 1} u(n)x^n = \sum_{k \ge 0} 2^k \sum_{n \ge 3k} \binom{n-k}{2k} x^n = \sum_{k \ge 0} 2^k \frac{x^{3k}}{(1-x)^{2k+1}},$$

$$\sum_{n \ge 1} v(n)x^n = \sum_{k \ge 1} 2^{k-1} \sum_{n \ge 3k} \binom{n-k-1}{2k-1} x^n = \sum_{k \ge 1} 2^{k-1} \frac{x^{3k}}{(1-x)^{2k}}.$$

Thus f is the sum of two geometric series. We obtain

$$\begin{aligned} f(x) &= \frac{1}{1-x} \cdot \frac{(1-x)^2}{(1-x)^2 - 2x^3} + \frac{1}{2} \cdot \frac{2x^3}{(1-x)^2 - 2x^3} \\ &= \frac{1 - x + x^3}{(1-2x)(1+x^2)} = \frac{1}{1+x^2} + \frac{x}{1-2x} = \sum_{n \ge 0} (-1)^n x^{2n} + \sum_{n \ge 0} 2^n x^{n+1}. \end{aligned}$$

Solution III by Donald E. Knuth, Stanford University, Stanford, CA. Let $t(n,k) = 2^k \binom{n-k}{2k} n/(n-k)$. The algorithm of Gosper and Zeilberger (explained in *Concrete Mathematics*, 2nd edition, by R. L. Graham, D. E. Knuth, and O. Patashnik (Addison-Wesley, 1994)) quickly yields

$$2t(n,k) - t(n+1,k) + 2t(n+2,k) - t(n+3,k) = 2t(n,k) - 2t(n,k-1).$$

Summing over $0 \le k < n$ yields $2s(n) - s(n+1) + 2s(n+2) - s(n+3) = 0$ when $n > 1$, since the right-hand side telescopes to $2t(n, n-1) - 2t(n, -1) = 0 - 0$ and since $t(n,k) = 0$ for $n/3 < k < n$. This is the recurrence (1), and the solution follows.

Editorial comment. Michael Hoffman considered a more general summation. When n, q, and $p + q$ are positive integers, let $F_{n,p,q}(t) = \sum_{0 \le k \le n/(p+q)} t^k \frac{n}{n-pk} \binom{n-pk}{qk}$. With $F_{0,p,q}(t) = 1$, Hoffman proved that

$$\sum_{n=0}^{\infty} F_{n,p,q}(t)x^n = \frac{(1-x)^{q-1} + (p/q)tx^{p+q}}{(1-x)^q - tx^{p+q}}.$$

He thus obtained $F_{n,-1,2}(-2) = \cos(n\pi/2)$. John Henry Steelman obtained $F_{n,1,1}(1) = L_n$, where the Lucas numbers $\{L_n\}$ are defined by the Fibonacci recurrence with $L_1 = 1$ and $L_2 = 3$.

Solved also by J. Anglesio (France), J. C. Binz (Switzerland), J. T. Bruening, D. Callan, E. R. Canfield, R. J. Chapman (U. K.), D. A. Darling, S. B. Ekhad, M. Hoffman, L. N. Howard, W. P. Johnson, D. Krug, J. H. Lindsey II, O. P. Lossers (The Netherlands), K. McInturff, I. Nemes & P. Paule (Austria), C. R. Pranesachar (India), R. Richberg (Germany), E. Schmeichel, J. H. Steelman, A. N. 't Woord (The Netherlands), M. Vowe (Switzerland), NSA Problems Group, and the proposer.

REVIEWS

Edited by **Underwood Dudley**

Mathematics Department, De Pauw University, Greencastle, IN 46135

Mathematical Circles (Russian Experience) by Dmitri Fomin, Sergey Genkin, and Ilia Itenberg. Translated from Russian by Mark Saul. American Mathematical Society, 1996, 272 pp, $29.00

Reviewed by **Andre Toom**

I shall call this book *Circles* for short. *Circles* is written by three Russian mathematicians with whom I have much in common. For many years we participated in similar activities including organization of mathematical competitions and teaching informal classes called "circles" in Russia. Six years ago I moved to the United States and worked at several universities. Based on this experience I want to reflect on how *Circles* can be used in America.

First of all, it is a rich collection of good problems. In addition there are useful notes for teachers. I think that Appendix A, which describes several types of mathematical contests, will be especially interesting for those who are dealing with all forms of cooperative learning. To get some taste of the book, let us look at a few problems.

Problem 1 on page 1
A number of bacteria are placed in a glass. One second later each bacterium divides in two, the next second each of the resulting bacteria divides in two again, et cetera. After one minute the glass is full. When was the glass half-full?

Some students propose a half-minute as the answer, implicitly assuming that the growth is linear. This problem shows in a dramatic way how different is exponential growth from linear.

Problem 10 on page 172
An evil king wrote three secret two-digit numbers a, b, and c. A handsome prince must name three numbers X, Y, and Z, after which the king will tell him the sum $aX + bY + cZ$. The prince must then name all three of the King's numbers. Otherwise he will be executed. Help him out of this dangerous situation.

It may seem at first that it is impossible to determine three variables from one equation and that the prince is doomed, but in fact he can save himself. The solution is based on a fruitful idea that can be used to introduce the students into the basics of information theory.

Circles may be very useful wherever there are classes devoted to solving non-standard problems. There may be more such classes than we are aware of at some selective schools. But Russian circles were not selective in any formal sense; in fact, anybody might drop in. So I did about forty years ago: I simply took a trolleybus, went to the old university campus in downtown Moscow and started to

attend informal classes taught by students of Moscow University. I submitted no formal application, paid nothing, and got no grades, but there I became a professional mathematician. Most classes discussed olympiad-style problems, but our teacher, Sasha Olevskii, was keen on epsilon-delta arguments of mathematical analysis, and for me it was just the right thing. Later I taught circles for decades, and now I meet my former students at various places, including Boston and Tel Aviv.

However, I think that the possible use of *Circles* is wider. I cannot imagine mathematical education without teaching students to solve mathematical problems. I am astonished when educators discuss some special "problem-solving approach" to teaching mathematics as if the opposite approach ever made sense. Universities offer special "problem-solving" courses. Does this mean that other courses do not teach to solve problems? Anyway, "problem-solving" courses should be welcomed, and *Circles* deserves a very prominent place among candidates for textbooks for such courses. In addition, many universities offer courses in mathematics without specifying exactly what these courses are about. When I teach such courses, I try to teach my students to solve as many problems as possible. It was very difficult for me to find a textbook for that. Now I have found at least one: *Circles*.

At this point I must admit some inconsistency. I have to agree with the translator's disclaimer: "This is not a textbook." At the same time I am going to use *Circles* as a textbook. The point is that it is an unusual textbook. It is unusual in several respects, but let us concentrate on the following. First, most problems in *Circles* need a rigorous approach. Some of them explicitly require proofs, while others ask questions whose answers involve argumentation. Second, most problems in most textbooks can be solved in a way that is explained in advance. Many such problems have identical mathematical structure presented in different "real world" guises. All this is absent in *Circles*. Most problems need a new idea although problems are grouped and some general theory is explained.

Now we approach a very important notion: *transfer of training*. Let us imagine the set of all possible problems in a branch of mathematics as a metric space. Every particular problem is a point in this space and similar problems are close to each other. By discussing a problem with our students, we cover a sphere with the center at this problem and radius equal to our students' ability to transfer their training from this problem to similar problems. Our purpose is to cover the greatest possible space with these spheres. Is transfer of training possible? For the authors of *Circles*, for me, and for all who have ever taught in this style the answer is evident: "yes, of course, transfer of training is possible and it is closely related to another precious human ability: generalization."

But some American educators have questioned the importance or the very possibility of transfer of training. Instead they proposed to solve only such problems in class that people face in everyday life. These educators have sent many messages to the effect that students are not interested in problems that have no immediate real-world use. As a result, some students' ability for generalization and transfer of training is almost completely atrophied. As soon as a problem on a test slightly deviates from problems solved in advance, they complain, "We did not solve such problems." Whichever course these students take, they learn to solve only those particular problems that the teacher chose to explain. Their spheres have radii that are close to zero and their total measure is also close to zero. It is a safe bet that the problems they have to solve after graduation practically never coincide with those few they have learned to solve in class.

All this does not mean that the manner in which problems are formulated is unimportant. It is very important and we may reproach *Circles* for some neglect in this respect. Let us illustrate these ideas by some examples.

Problem 11 on page 53
A woodman's hut is in the interior of a peninsula which has the form of an acute angle. The woodman must leave his hut, walk to one shore of the peninsula, then to the other shore, then return home. How should he choose the shortest such path?

This is a very useful problem. It introduces students to the important realm of geometrical transformations. It also gives the teacher a chance to speak about relations between mathematics and physics because light also "chooses" the shortest path. Regrettably, no motivation is given why the woodman should walk in such a strange way. In Steinhaus's book a similar problem is formulated more vividly: "An Arab wishes to return to his tent, but on the way he wants to feed his horse and draw water from a river." [2, pp. 111–112]. Note that my criticism has nothing to do with the silly requirements of straightforward "real-world" relevance. The real relevance is through theory, as usual. Another example: in problem 22 on page 3 there is a strange river that makes a 90° turn. In Kordemsky's book [1, problem 126 on p. 54] it was a moat surrounding a fort, which was more plausible and romantic.

Since I started to criticize *Circles*, let me continue to do it from my point of view as a teacher. *Circles* does not mention irrational numbers at all, not even the famous proof by contradiction that no square of a rational number equals 2. It is very easy to collect a series of problems in this vein. (This is what I do in my classes.) Why didn't the authors of *Circles* do it?

Some other important facts are present but not emphasized. For example, it takes attentive reading to find problem 53 on page 29, which is included into a section called "A few more problems" as if it were something optional: *Prove that there are infinitely many prime numbers*. Nothing is said about the importance of this fundamental fact.

On page 177 the authors explain one method to invent problems, which they call "inequalities à la Leningrad." A typical example, problem 1 on page 175: *Which number is greater*: 31^{11} *or* 17^{14}? Problems of this sort seem to have been a useful contribution to olympiads in the past, but I am afraid that the presence of calculators kills most of them. On the other hand, look at problem 43 on page 161: *If all the sides of a triangle are longer than* 1000 *inches, can its area be less than* 1 *square inch*? This problem and several similar ones were invented in the Moscow-based School by Correspondence, and these problems remain usable in presence of any technology.

Problem 10 on page 23 assumes acquaintance with divisibility tests for 3 and 9, which are introduced only in problem 31 on page 99. The same seems to be true for problem 76 on page 72; I could solve it only using divisibility tests. There is some confusion with problem 65 on page 71. The letter M is misplaced and K is absent in figure 123 on page 163.

I admit another inconsistency about students' age. *Circles* is addressed to 12–14 year old children. This contrasts with the way I am going to use *Circles*. All of my students are intelligent and motivated, but they are around twenty years old or more. Sometimes I wonder what have they been doing for so many years. Now they have to think about graduation and making a career. They are pressed by time and

money concerns, some have to do odd jobs, some have children. Of course, all this interferes with their study. It would be much better for them to have started to solve non-trivial problems years ago.

This inconsistency is especially noticeable when one reads the initial Chapter Zero, intended for students of ages 10–11. "The problems of this chapter have virtually no mathematical content," naively claim the authors. Actually the problems of this chapter require the most fundamental ability—the ability for abstract thinking. This ability should by no means be taken for granted; it develops as a result of careful and well-thought schooling. Let us remember that most teachers of circles in Russia were not professional teachers. In most cases they were university students, inspired but unexperienced. Their teaching was successful due to the sound preparation provided by the national educational system. In my young years the Russian educational administration seemed very stupid to me, but now I see how efficient actually it was. Don't ask me how does this square with the tyrannical Soviet rule and the ailing Russian economy because I don't have all the answers.

Some people ask whether there are competent enthusiasts in America who could and would teach classes of creative problem solving. The answer is "yes, of course," but this is not the right question to ask. The right question is whether the educational system can teach the basics of mathematics so that children will be able to attend such classes.

REFERENCES

1. Boris A. Kordemsky, *The Moscow Puzzles*, 359 *Mathematical Recreations*. Dover Publications, 1992.
2. H. Steinhaus, *Mathematical Snapshots*. New York, Oxford University Press, 1969.

Department of Mathematics
University of the Incarnate Word
4301 Broadway
San Antonio, TX 78209
toom@the-college.iwctx.edu

Vita Mathematica: Historical Research and Integration with Teaching. Edited by Ronald Calinger. Washington, D.C., Mathematical Association of America, 1996, 350 pp., $34.95.

Reviewed by **Hardy Grant**

That mathematics has an "image problem" with large segments of the general public has been obvious for a long time. The stock indictment is all too familiar, for it has echoed repeatedly in the cultural creations that reflect and shape our collective values. Since Aristophanes, mathematics and its devotees have been lampooned, gently or savagely, as narrow, austere, mechanical, passionless, excessively cerebral, cold. Examples abound. Jane Austen's Emma (1816) considers at one point that the romantic events unfolding around her must impinge on "the coldest heart and the steadiest brain," on a linguist, a grammarian, "even [!] a mathematician." The "Master Mathematician" in Oscar Wilde's *The Happy Prince* (1888) "frowned and looked very severe, for he did not approve of children dreaming." Have perceptions changed in our own time? A few months ago, the

parents in the enjoyable comic strip *Sally Forth*, weighing possible summer destinations for their 9-year-old, threatened her with "Algebra Camp" ("just outside Geek City") as the absolute antithesis of a place where a kid could hope to have some *fun*. Readers of the MONTHLY could multiply these instances at will. The problem runs very deep, and does not look like going away any time soon.

Is there in fact a cure? Countless adults blame their distaste for mathematics on bad teaching, so the schools would seem the natural place to start. But what to do there? For many the best hope remains the use of history in the mathematics classroom as a "humanizing" corrective. The Danish educator Torkil Heiede goes so far as to say, in his contribution to the volume under review, that to teach mathematics "without its history" is to teach it "as if it were dead." The book's origin reflects the now widespread sharing and institutionalization of such sentiments. Many of the papers assembled here began as talks given in the summer of 1992 at either the Quadrennial Meeting of the International Study Group on the Relations between History and Pedagogy of Mathematics ("HPM"), at Toronto, or the Seventh International Congress on Mathematical Education ("ICME"), at Quebec City. According to Ronald Calinger, the volume's editor, other papers are the fruits of a call sent out by him to "historians of mathematics." The book's subtitle, "Historical Research and Integration with Teaching," captures well enough such focus and unity as this wildly diverse collection may be said to possess. Calinger identifies the primary audience as "mathematics teachers, research mathematicians, historians of mathematics, and historians of science" (p. vii).

Predictably, these thirty articles vary enormously in subject matter, in level of presentation, and in putative audience. Someone, presumably the editor, has gamely tried to partition them according to theme, but the subsets' boundaries are rather slippery, and indeed I would argue that perhaps half a dozen contributions have ended up in the wrong boxes. I shall adopt a somewhat different grouping in what follows. I should add that any attempt to summarize so many papers in so brief a compass must be lamentably selective and superficial, and due apology is hereby extended both to their authors and to potential readers.

A preliminary word about the book's "production values" may be in order. The index is good, and in many articles the bibliography alone is worth the price of admission. Many marvellous illustrations have been culled from inaccessible sources. The writing is mostly very competent, if occasionally clumsy and seldom inspired. Surprisingly or not, some of the authors for whom English is presumably a learned language write it better than do some presumably native users. Alas, these pages will not reassure those readers (I am one) who fear that in our time professional standards of editing and proofreading are in steep decline. Thus (to offer a sampling that is very far from exhaustive) "data" is stubbornly treated as singular (pp. 92, 335), and "phenomena" too for good measure (p. 332); *two* verbs on p. 94 disagree with their respective subjects; several references of a paper to itself have been left as "p. ???"; and computer gremlins have at several places been allowed to substitute commas for occurrences of "é." (One might have thought that "Dieudonn," (pp. 9, 10) would rouse the sleepiest proofreader; but apparently not.) It is only fair to add that the passages involving technical mathematics seem to have been treated with commendable care.

Not much in these pages will jolt readers out of their seats, or spark revolutions. Few seeds of potential controversy are here sown. But one paper, David Rowe's fine survey of trends in the historiography of mathematics, deals at some length with old disagreements that still smolder; I shall come back to that paper, and those issues, toward the end of the review. The author here whom I should most

enjoy debating face to face is Ubiratan D'Ambrosio, who offers a spirited defence of an "ethnomathematics" conceived so broadly as to coincide, so far as I can see, with all of cultural anthropology ("the study of techniques developed in different cultures for explaining, understanding, and coping with their physical and socio-cultural environments"). Regrettably, a definition so vague and idiosyncratic seems likely to take the paper out of the mainstream discussion of its ostensible subject.

One of the book's great delights is its several ventures along relatively untrodden byways of our mathematical heritage. Complete with charming cartoons (by the author?), Beatrice Lumpkin's article cleverly interweaves the history of the rule of false position (which she traces to ancient Egypt) and the career of the outstanding mathematics educator Benjamin Banneker (d. 1806). Peggy Kidwell tells the absorbing story of the various objects, especially geometric models, that saw service as teaching aids in 19th-century schools and universities. Kidwell links these long-ago educators' enthusiasm for such objects to their belief that "it was more useful to show pupils the properties of surfaces and solids using models than to offer formal proofs of those properties"—a notion that has not lost its validity. Karen Dee Michalowicz rescues from obscurity Mary Everest Boole (the mountain is named for her uncle), wife of George Boole and a distinguished writer on, and teacher of, mathematics in her own right. Those who still doubt that women have "come a long way" in our subject may like to ponder a despairing remark made by Mary's father in 1842, when she was ten. If she could go to university, he said (which she could not), she "would carry everything before her ... But what could a girl do learning mathematics?" This article includes two-plus pages of short quotations from Mary Boole's writings on the art of teaching, a little treasury that I wouldn't trade for ten years' worth of "aims and objectives" from my local Ministry of Education.

Vita Mathematica contains a number of papers on "straight" history of mathematics, without overt or obvious pedagogical application. Jens Høyrup traces from old Babylonia to the Renaissance the career of a single problem—to find the side of a square from the sum of its perimeter and area; this, he says, turns out to belong to a non-scholarly tradition of practical geometry, whose interaction with the contemporary "literate" mathematical culture he discusses at length. Wilbur Knorr's theme is the ancient practice of the "method of indivisibles" and its influence in the age of Cavalieri. Knorr argues that an "indivisibilist heuristic" predated Archimedes, and that early-modern versions represent "a reconstruction of the lost heuristic by geometers who were impatient over the demands of formal demonstration." These two papers are substantial monographs, with the detail and rigor of argument and documentation that we expect from their authors. Shorter and slighter, but perhaps on that account better suited to "integration with teaching," are an overview of traditional Chinese mathematics by Frank Swetz, a discussion of combinatorics and induction in medieval Hebrew and Islamic mathematics by Victor Katz, and Barnabas Hughes' description of the earliest correct algebraic solution of cubic equations, by Master Dardi of Pisa (c. 1350). All three of these essays are solid and useful. Zarko Dadić well summarizes the work of the Croatian scientist Marin Getaldić (1568–1626). Getaldić wrote *inter alia* a treatise *De resolutione et compositione mathematica*—which is to say, the use in mathematics of the method of "analysis and synthesis." Some day that volume, and Dadić's paper here, will be sought out by the author of one of the great unwritten books of the western intellectual tradition, a history of analysis and synthesis in their two millennia of life, not just in mathematics but in philosophy and early-modern science as well.

The other purely historical essays in *Vita Mathematica* take the reader much nearer to our own time. Roger Cooke gives an admirable account of Sof'ya Kovalevskaya's "mathematical legacy," in particular her "discovery of a physical configuration for which the equations of motion of a rigid body about a fixed point under the influence of gravity can be integrated in closed analytic form." William Aspray, Andrew Goldstein, and Bernard Williams entertainingly recount the rise of theoretical computer science and engineering, in both their intellectual and social aspects but with emphasis on the latter, specifically the supporting role of the National Science Foundation.

Other papers treat of the history of mathematics education. Hans Niels Jahnke, in one of the book's best essays, relates the complicated 19th-century history of "algebraic analysis"—the most familiar manifestation is probably Lagrange's recasting of the calculus in terms of power series—and explains how its vogue had the remarkable effect of excluding infinitesimal methods from the curricula of German gymnasia for several decades. Jahnke argues that pedagogically this was no bad thing, for it entailed a stress on "concrete" problem-solving as opposed to blind application of algorithms. Susann Hensel's topic is the mathematical education of engineers in late-19th-century Germany. As she says, some of the questions then under debate are hardy perennials: how "pure" and rigorous should the mathematics in "service" courses be, and should they be taught only by mathematicians? Ronald Calinger focusses on the early years (1861 ff.) of the famous mathematics seminar at Berlin; the midwives at its birth were Kummer and Weierstrass. Calinger weaves into his tale a good deal of the mathematics of these two giants and of their contemporaries.

The papers in yet another group get closer to the volume's professed goal of integrating history with teaching. Fred Rickey makes the case in general terms, describes particular techniques in his own practice, and calls attention to such resources as the HPM newsletter and the MAA's thriving e-mail discussion group on the history of mathematics. (Rickey is too modest to mention here that the latter is his own creation.) Torkil Heiede relays the welcome news that the Danish government has *mandated* the use of the history of mathematics in "upper grades." One longs to know more: who had the ears of the bureaucrats, and what arguments were used that might be transplanted to less progressive jurisdictions?

A repeated theme in these pages is the value of history in promoting a view of mathematics as a *process*, rather than merely a product, of human striving and discovery. This contrast explicitly guides Evelyn Barbin's persuasive advocacy, with historical examples, of a problems-oriented approach to teaching. The benefits, she says, include a better understanding and tolerance of pupils' errors. Similarly Manfred Kronfeller, after providing a short history of the function concept, draws pedagogical lessons that include the realization that student errors may actually mimic those of the great masters. Indeed, Kronfeller says, in teaching any set of ideas one should follow as much as possible their historical evolution, for one can "assume" that aspects elucidated earlier in history should be easier for pupils to grasp. Some empirical evidence in the same direction is offered by Peter Bero, who conducted in Slovakia a survey of young people's perceptions of the continuum. (The nodding proofreaders here provide the funniest of the book's innumerable typos: the age range of Bero's respondents is said (p. 304) to be "1 to 18." So what *does* the playpen set make of Zeno?) Bero reports that these elementary-school and gymnasium students conceive the continuum in ways "often similar to those displayed by ancient Greek mathematicians." One thinks of the biologists' catchy

old saying that "ontogeny recapitulates phylogeny"—the individual's development retraces the evolution of her species.

Other authors seek to base teaching on the direct use or creative reworking of primary sources. Richard Laubenbacher and David Pengelley describe an upper-level honors course that presents students with "mathematical masterpieces," ranging in time from Archimedes to John Conway, and asks them to function as "critics" in the sense in which this term is used in the arts. Israel Kleiner outlines a course, for *teachers*, built around the use of quotations about mathematics, supplemented by a "very concise" chronology. Very instructive here are a number of *pairs* of mutually opposing quotations, which give a vivid sense of the complexity of issues and of their inherent drama. This potential for drama is taken to its natural conclusion by Gavin Hitchcock, who has written stageable dialogue in which figures from the history of mathematics debate their respective stances. In the first of these playlets Simon Stevin touts irrational numbers against the qualms of Michael Stifel; in the second, set in 1827, George Peacock and Augustus de Morgan try to make William Frend (himself a mathematician) accept multiple and negative roots of equations. Given the right actors and audience, these exchanges could "come off the page" with real theatrical effectiveness.

Several authors present case studies that extract pedagogical morals from specific historical episodes. John Fauvel's point of departure is a wonderfully curious "tree" diagram, from a book of 1808, which shows graphically the various factors that led to the abolition of the slave trade. Fauvel's agenda here is threefold. He argues that (i) "graphical representation and modelling" have been neglected by historians of mathematics, (ii) they have also been "devalued" by teachers, in comparison with "prose text or algebraic symbolism," and (iii) they can and should be used to "empower" students, that is, to encourage in students the "knowledge and belief that they can use and create mathematics to influence their way in the world." Marie Françoise Jozeau and Michèle Grégoire describe the meridian measurements in France (1792–99) that led to the first definition of the meter, and tell also (a bit cursorily, to my regret) of an imaginative reenactment, near Paris, of part of that labor by a group of modern high school students. Jim Tattersall sketches the history of attempts to estimate the total number of people who ever lived, and proposes a classroom exercise to the same end. This piece revives an ancient wheeze which is the best joke in a book not brimming with humor. A class of students was asked how they knew they were going to die, and one replied brightly that "it was because so far most people have."

Several other papers (in addition to Tattersall's) draw their case studies from the history of the calculus. Judy Grabiner summarizes the sharply contrasting approaches taken to the calculus by Maclaurin (geometric) and by Lagrange (algebraic), respectively. That is a familiar story, but Grabiner goes on to trace the difference to differing cultural influences, and sets out the lesson for teachers: the tendency of students to adopt diverse approaches to mathematical problems is both natural and legitimate, especially where "non-traditional" backgrounds are involved. Martin Flashman suggests that instructors give an extra historical dimension to the textbook account of the Fundamental Theorem of the Calculus by presenting in detail the wholly geometric version proved by Isaac Barrow just before the watershed work of Newton and Leibniz. Man-Keung Siu considers "integration in finite terms" from its first rigorous treatment by Liouville (1830s) to the modern classroom. He gives splendid expositions both of the mathematics itself —which is far from elementary—and of its history, and in both aspects he

maintains an exemplary balance between superficiality on the one hand and excessive detail on the other. And Siu is equally good on the pedagogical issues. He has a fine sense of the kind of question that better students are likely to raise, and he specifically addresses the impact of computers on the teaching of his chosen topic. For my money his paper is another of the volume's highlights.

I come finally to the excellent essay, already cited above, in which David Rowe depicts "new trends and old images" in the historiography of mathematics. In particular he discusses the notorious debate that climaxed in the late 70s over the "geometrical algebra" commonly credited to the Greeks. Many results stated and proved in geometrical language by Euclid can easily be "translated" into elementary algebraic identities; but does such restatement distort the Greeks' own point of view? So argued the historian Sabetai Unguru, whose views then evoked rebuttals from mathematicians of the stature of André Weil, Hans Freudenthal, and B. L. van der Waerden. Rowe expounds this particular dispute with applaudable fairness, and then sets it in a larger context, centering his discussion around some provocative views voiced by the same André Weil in a famous lecture in 1978. Weil declared on that occasion that *for mathematicians* the "first use" of the subject's history is "to put or keep before our eyes 'illustrious examples' of first-rate mathematical work." Therefore "the craft of mathematical history can best be practiced by those of us who are or have been active mathematicians or at least who are in contact with active mathematicians." Moreover, Weil urged, it is appropriate, indeed necessary, to interpret the mathematical ideas of past cultures in modern terms. For example, "it is impossible for us to analyze properly the contents of Books V and VII of Euclid [on ratios of magnitudes and on number theory, respectively] without the concept of group and even that of groups of operators, since the ratios of magnitudes are treated as a multiplicative group operating on the additive group of the magnitudes themselves."

The tendency exhibited in this last quotation is probably more seductive in mathematics than in any other subject. For in mathematics, as the same quotation shows, many ancient ideas can be fitted with perilous ease into modern abstract frameworks to which they are wholly foreign in conception and in spirit. Perhaps then it was no great surprise to read recently a hint that mathematics may have acquired among outsiders a certain reputation for this kind of thing. In the obituary of Joseph Needham that she wrote for the June 1996 issue of *Isis*, Francesca Bray tossed out the aside that "projecting modern meanings onto ancient terms [is] a practice that often passes unnoticed, I am told, in the history of Western mathematics." "Often," indeed; for the temptation is chronic.

But not, in, fact, "unnoticed." It was just this tendency that Sabetai Unguru protested back in the 70s, and now David Rowe takes up the cause. Still addressing himself specifically to the views of André Weil, quoted above, Rowe explains with great civility what "disturbs" historians about viewing ancient ideas through modern lenses. He associates Weil's vision of history with a Platonist philosophy, which makes mathematical truths independent of time and of historical milieu. As he says, this orientation disposes one to "present the development of mathematical ideas as a steadily unfolding search for Platonic truths that transcend the particular cultural contexts in which these ideas arose" (p. 10). But he raises at once, in the completion of the sentence just quoted, the obvious objection, that one can write history in this way "only by discounting the rich variety of meanings that accompanied" those ideas in their actual concrete settings. With this objection I agree entirely—but I would push the argument a bit further. There is a facet of the issue that most discussions mention scarcely if at all.

I suppose that those who credit the ancients with this or that modern idea think that the ascription amounts to generous praise. For all their limitations, the argument goes, these forerunners had the Right Stuff, they were spiritually so close to us that they seem (in Littlewood's memorable phrase) like "fellows of another college"; and wouldn't they rejoice in our good opinion of them? Well, maybe so. It might just be, however, that the mathematicians of bygone ages would also, or even *rather*, wish to be studied and judged *for what they were*, in their own proud individuality and distinctiveness. In the writing of history within the traditional academic discipline of that name, the goal of an empathetic portrayal of the past *on its own terms* is now a commonplace. This is not at all to say, of course, that anticipations of, influences on, the present are not worth notice. But the aim is the polar opposite of the deliberate backward projection of modern modes of thought and feeling. Rather, the quest is for what William Blake would call the "minute particulars," the specific and defining uniqueness, of each time and place and people of the past.

Historiography so conceived has at its very heart a *moral* imperative, a scrupulous respect for, and honoring of, those who went before. At its highest, it is—dare one say?—like an act of love, which would know and possess yet ultimately leave singular and autonomous. That makes its practice at once a noble and a formidably demanding enterprise. George Steiner once wrote, in a different but parallel context, that "there can be no other thanks [to the creators of our cultural heritage] than extreme precision, than the patient, provisional, always inadequate attempt to get each case right." That challenge is present no less in the historiography of mathematics than elsewhere, despite the (for some) supposedly eternal character of the truths unveiled. Indeed, in the last analysis this timelessness is a red herring, for the most passionate Platonist must concede that mathematical *discovery* is the work of individual minds in historically conditioned settings.

So who should write the history of that discovery? Where the object of study is the technical progress made in the 20th century, André Weil's claim that only mathematicians need apply seems incontestable. In most fields of current mathematics, the dynamic of the development is so overwhelmingly "internal," and the barriers of specialist knowledge so forbidding, that the prospective historian cannot hope to succeed without something very close to the active researcher's intimate grasp of, and "feel" for, the subject. The scarcity of people qualified in this sense must be the prime reason for the deplorable under-representation of the 20th century in the journals and conference programs devoted to serious history of mathematics.

But if the historiography of 20th-century mathematics must be left to the practitioners, the past—even the relatively recently past—is a different case. Roger Cooke says acutely in this volume (p. 177) that, thanks above all to the massive modern trend toward abstraction and generalization,

> To reconstruct precisely the present state of any nineteenth-century mathematical topic is in a sense impossible. No mathematical problem is understood exactly as it was understood ... a century ago.

Then how much wider still the gulf between us and the still more distant past! The farther back in time, the more foreign and elusive must be the "mindset" that the historian would seek to know. Moreover, and crucially, increasing remoteness from the present increases also the role of "external" factors in mathematical activity—and so diminishes, in proportion, the place of purely mathematical skills

and knowledge in the aspiring historian's stock in trade. Where ancient mathematics is in question, the historian needs little if any technical grasp beyond the very elementary levels (quadratic equations or whatever) achieved by the civilization under her gaze. The resources that must sustain her are in other directions entirely. The importance of the relevant linguistic competence is obvious. The whole of the surrounding cultural and social matrix is germane and must be mastered. Perhaps the insights of modern anthropology can be brought to bear, as by Geoffrey Lloyd in his superb studies of ancient Greek science. Above all, the successful explorer of mathematics' distant past must bring the precious gifts of imagination and of empathy that alone give any hope of access to alien minds. In this terribly difficult undertaking the research mathematician as such has absolutely no privileged status, no claim whatever to special authority.

The good things in the book under review, and there are many, make their own valuable contributions to the history of mathematics and to the creative use of that history in our classrooms. Thus they may serve also toward fulfilling the hope, articulated here by Fred Rickey, of educating "a general population with a much better feel for what mathematicians do and why it is important." In that urgent task *Vita Mathematica* will not set the world on fire, but it should light some candles; and every bit helps.

539 Highland Avenue
Ottawa, Ontario K2A 2J8 Canada
hgrant@freenet.carleton.ca

Contributed by Russ Hood, Rio Linda, CA

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of
Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1–4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, T*(13–14: 2). *Principles and Practice of Mathematics.* COMAP. Springer-Verlag, 1997, xi + 686 pp, $64.95. [ISBN 0-387-94612-8] A thoughtful alternative to calculus as an entry point to college mathematics; same level and prerequisites as first-year calculus. Prepares students for more advanced studies in mathematics and complements the calculus sequence. Topics: sequences and difference equations, vectors, some analytic geometry, linear programming, basic linear algebra, combinatorics, graphs and algorithms, logic, probability and decision theory, symmetry and permutation groups, coding. Extensive applications. Gives students both specific tools and a sense of the breadth of mathematics used in scientific and industrial settings. LB

Recreational Mathematics, S(13–16), P, L**.** *From Erdös to Kiev: Problems of Olympiad Caliber.* Ross Honsberger. Dolciani Math. Expos., No. 17. MAA, 1996, xii + 257 pp, $31 (P). [ISBN 0-88385-324-8] Only a master expositor could take these problems, mostly from the 1987 and 1988 volumes of *Crux Mathematicorum*, and present them as works of art with direct appeal to general readers. The author is quick to draw attention to subtleties that make the problem interesting; his solutions are exquisitely clear and easy to read, instructive and pleasurable. LCL

Recreational Mathematics, S, L*.** *Five Hundred Mathematical Challenges.* Edward J. Barbeau, Murray S. Klamkin, William O.J. Moser. Spectrum Ser. MAA, 1995, xi + 227 pp, $29.50 (P). [ISBN 0-88385-519-4] These problems, first published in a series of booklets almost twenty years ago, span all areas of high school mathematics (pre-calculus), and range in difficulty and sophistication from easy puzzles to problems at the Olympiad level. A useful appendix summarizes essential knowledge (combinatorics, arithmetic, algebra, inequalities, geometry and trigonometry, analysis). A short index classifies problems by subject. LCL

Recreational Mathematics, S*(16–18), P, L.** *Contests in Higher Mathematics: Miklós Schweitzer Competitions 1962–1991.* Ed: Gábor J. Székely. Problem Books in Math. Springer-Verlag, 1996, vii + 569 pp, $59. [ISBN 0-387-94588-1] Questions, with detailed solutions, from this unique and prestigious Hungarian mathematics take-home exam. (A big step beyond the Putnam; students may use materials available in libraries or homes, and have ten days to prepare their solutions). Topics include algebra, combinatorics, theory of functions, geometry, measure theory, number theory, operators, probability theory, sequences and series, topology, and set theory. LCL

Recreational Mathematics, S, L**.** *Leningrad Mathematical Olympiads, 1987–1991.* Dmitry Fomin, Alexey Kirichenko. Contests in Math., V. 1. MathPro Pr, 1994, xix + 197 pp, $24 (P). [ISBN 0-9626401-4-X] An outstanding collection of inviting challenge problems graded by level of difficulty from middle school to high school. Many are suitable for undergraduate "Problems-of-the-Week" recreations (arithmetic, algebra, combinatorics, dis-

crete mathematics, geometry, functions, beginning analysis). LCL

Recreational Mathematics, S, L**.** *The Universe in a Handkerchief.* Martin Gardner. Copernicus (Springer-Verlag), 1996, x + 158 pp, $19. [ISBN 0-387-94673-X] A friendly, informative introduction and guide to the ingenious and intriguing games, puzzles, and word plays of Lewis Carroll. LCL

Recreational Mathematics, S, L. *Rediscovered Lewis Carroll Puzzles.* Ed: Edward Wakeling. Dover, 1995, xiii + 79 pp, $4.95 (P). [ISBN 0-486-28861-7] Forty-two recreational puzzles and games (with solutions) used by Lewis Carroll to entertain friends, colleagues, and "children of ages five to ninety-five." LCL

Recreational Mathematics, S, L. *ARML–NYSML Contests, 1989–1994.* Lawrence Zimmerman, Gilbert Kessler. Contests in Math., V. 2. MathPro Pr, 1995, xviii + 189 pp, $19.95 (P). [ISBN 0-9626401-6-6] A sequel to the *ARML–NYSML Contest Book 1973–1985* published by NCTM in 1987. These contest problems for high school students range from short answer questions to challenging multi-part problems requiring in-depth analysis and original thinking. All problems are original; instructive solutions. LCL

Recreational Mathematics, S. *Collection of Problems on Smarandache Notions.* Charles Ashbacher. Erhus Univ Pr, 1996, 73 pp, $8.25 (P). [ISBN 1-879585-50-2] Collection of open problems in number theory, mostly of the following sort: Which elements of a given sequence S have property P? For example: Which triangular numbers, with digits x_i, can be "partitioned," for some $1 \leq k < m < n$, into the form $x_1x_2 \dots x_kx_{k+1} \dots x_mx_{m+1} \dots x_n$ so that $x_1x_2 \dots x_k + x_{k+1} \dots x_m = x_{m+1} \dots x_n$? LCL

History, L. *The Way I Remember It.* Walter Rudin. History of Math., V. 12. AMS, 1997, ix + 191 pp, $29. [ISBN 0-8218-0633-5] Walter Rudin's memoirs. Includes samples of his work. Written for non-analysts. LC

History, L*. *Poincaré and the Three Body Problem.* June Barrow-Green. History of Math., V. 11. AMS, 1997, xvi + 272 pp, $49. [ISBN 0-8218-0367-0] Account of Poincaré's memoir on the three-body problem from a mathematical and historical perspective. Also discusses earlier work by other mathematicians, reactions by Poincaré's contemporaries, and the memoir's influence on later work. LC

Foundations, T(16–17: 1), L. *Intermediate Set Theory.* F.R. Drake, D. Singh. Wiley, 1996, x + 234 pp, $29.95 (P). [ISBN 0-471-96496-4] Good intermediate level treatment of set theory including ZFC, first-order logic, cardinals and ordinals, axiom of choice, constructible sets and forcing, the standard paradoxes, and the development of mathematics within ZFC. RM

Combinatorics, P. *Matroid Theory.* Eds: Joseph E. Bonin, James G. Oxley, Brigitte Servatius. Contemp. Math., V. 197. AMS, 1996, xii + 418 pp, $72 (P). [ISBN 0-8218-0508-8] Proceedings of a 1995 AMS–IMS–SIAM Joint Summer Research Conference at the University of Washington.

Discrete Mathematics, S(14–15). *Exploring Discrete Mathematics With Maple.* Kenneth H. Rosen, *et al.* McGraw-Hill, 1997, viii + 392 pp, $20.97 (P). [ISBN 0-07-054128-0] Guide to Maple's discrete mathematics capabilities. Assumes no prior experience with Maple. Includes discussion and examples of built-in commands as well as procedures written for the book. LC

Number Theory, T(15–17: 2), S, P, L. *Number Theory: An Introduction.* Don Redmond. Pure & Appl. Math., V. 201. Marcel Dekker, 1996, xii + 749 pp, $175. [ISBN 0-8247-9696-9] Lots of exercises and problems, with optional computer investigations, accompany this expansive, instructive, and informative text. The first half contains material for an undergraduate introduction (primes, divisibility, congruences, quadratic residues, Diophantine equations); the second half demonstrates, in a self-contained way, how other areas of mathematics enter into the study of natural numbers (continued fractions, fractions), arithmetic functions, Bertrand's Postulate, the Chebychev Theorem, the prime number theorem, and algebraic number theory. LCL

Number Theory, P. *Sets of Multiples.* Richard R. Hall. Tracts in Math., V. 118. Cambridge Univ Pr, 1996, xvi + 264 pp, $59.95. [ISBN 0-521-40424-X] Asymptotic analysis, mostly via sieve methods, of number theory problems that can be phrased in terms of sets of integers for which any integer multiple of an element is also in the set. DB

Number Theory, S*(15–18), P*, L.** *Theory of Algebraic Integers.* Richard Dedekind. Transl: John Stillwell. Math. Lib. Cambridge Univ Pr, 1996, vii + 158 pp, $22.95 (P). [ISBN 0-521-56518-9] A translation of Dedekind's 1877 memoir explaining his ideal theory to a general mathematical audience. Extensive introduction by John Stillwell describes the history and problems of the study of algebraic integers. Accessible to good undergraduates. DB

Linear Algebra, S(14–16), L. *Linear Algebra:*

Challenging Problems for Students. Fuzhen Zhang. Stud. in Math. Sci. Johns Hopkins Univ Pr, 1996, ix + 174 pp, $14.95 (P); $35. [ISBN 0-8018-5459-8; 0-8018-5458-X] 200 problems with hints and solutions. Problems range from easy to difficult; some are standard problems found in any text. LC

Linear Algebra, T(15–18: 1). *Linear Algebra with Applications.* John T. Scheick. Intern. Ser. in Pure & Appl. Math. McGraw-Hill, 1997, xv + 432 pp, $57.75. [ISBN 0-07-055184-7] An extraordinary compilation of topics and applications from linear algebra both inside and outside mathematics. Well-written; highly mathematical. A worthy addition to your library. PF

Algebra, T(15–16: 2), L. *Fundamentals of Abstract Algebra.* D.S. Malik, John M. Mordeson, M.K. Sen. McGraw-Hill, 1997, xix + 636 pp, $64.63. [ISBN 0-07-040035-0] Standard topics (groups through Sylow theorems, solvability, nilpotence; rings and modules, including UFD's, Noetherian, Artinian, matrices; fields, Galois theory), plus sections on coding theory and Gröbner bases. Proofs very calculational; worked-out exercises after each section. RM

Algebra. *Groups as Galois Groups: An Introduction.* Helmut Völklein. Stud. in Adv. Math., V. 53. Cambridge Univ Pr, 1996, xvii + 248 pp, $49.95. [ISBN 0-521-56280-5] The first half develops the background (covering space theory, Riemann surfaces, number theory) to study the inverse Galois problem. The second half presents recent results (braid group actions, embedding problems, moduli spaces). TH

Real Analysis, T(17: 2). *Real Analysis.* Andrew M. Bruckner, Judith B. Bruckner, Brian S. Thomson. Prentice-Hall, 1997, xiv + 713 pp. [ISBN 0-13-458886-X] Includes material on measure theory, Banach spaces, Hilbert space, and pointwise convergence of Fourier series. Develops some of the theory in the (numerous) exercises. PG

Numerical Analysis, P. *The Mathematics of Numerical Analysis.* Eds: James Renegar, Michael Shub, Steve Smale. Lect. in Appl. Math., V. 32. AMS, 1996, xi + 927 pp, $125 (P). [ISBN 0-8218-0530-4] Proceedings of a 1995 AMS–SIAM Summer Seminar in Park City, Utah.

Algebraic Geometry, T(17: 2), P. *Combinatorial Convexity and Algebraic Geometry.* Günter Ewald. Grad. Texts in Math., V. 168. Springer-Verlag, 1996, xiv + 372 pp, $59. [ISBN 0-387-94755-8] Shows the relationship between combinatorial and algebraic geometry via torus embeddings. Topics: polytopes, polyhedral sets, toric varieties, sheaves, cohomology. Assumes only linear algebra and calculus. JD

Topology, T(18: 2), P. *Knots, Links, Braids and 3-Manifolds: An Introduction to the New Invariants in Low-Dimensional Topology.* V.V. Prasolov, A.B. Sossinsky. Transl. of Math. Mono., V. 154. AMS, 1997, viii + 239 pp, $99. [ISBN 0-8218-0588-6] Presents Jones and Vassiliev invariants in an elementary manner, but bulk of book is rigorous geometric treatment of Jones-Witten invariants. Includes Heegaard decompositions, surgery, Kirby calculus, branched coverings, but does not cover any physics. Exercises with solutions. JD

Operations Research, T(17–18), P. *Stochastic Programming Problems with Probability and Quantile Functions.* Andrey I. Kibzun, Yuri S. Kan. Intersci. Ser. in Systems & Optim. Wiley, 1996, xiii + 301 pp, $64.95. [ISBN 0-471-95815-8] Stochastic programming problems arise when optimizing a probability function that is the expectation of the indicator function on a set. Traditional optimization techniques fail because of the underlying discontinuity. Uses examples to establish the central problems, then presents a theoretical framework that allows the calculation of probability and quantile functions. MPR

Optimization, P. *Quasidifferentiability and Nonsmooth Modelling in Mechanics, Engineering and Economics.* Vladimir F. Dem'yanov, *et al.* Nonconvex Optim. & Its Applic., V. 10. Kluwer Academic, 1996, xvii + 348 pp, $169. [ISBN 0-7923-4093-0]

Probability, T(17: 1), P, L. *Probability: A Survey of the Mathematical Theory, Second Edition.* John W. Lamperti. Ser. in Prob. & Stat. Wiley, 1996, x + 189 pp, $39.95. [ISBN 0-471-15407-5] Designed for a second course in probability, assuming some background in measure theory (outlined in an appendix). Four chapters: Foundations; Laws of Large Numbers and Random Series; Limiting Distributions and the Central Limit Problem; the Brownian Motion Process. (*First Edition*, TR, October 1967; Extended Review, February 1969.) RSK

Stochastic Processes, T(16–17: 1), L. *The Analysis of Time Series: An Introduction, Fourth Edition.* C. Chatfield. Stat. Textbook Ser. Chapman & Hall, 1995, xiii + 241 pp, $34.95 (P). [ISBN 0-412-31820-2] Covers stationary, nonstationary, and bivariate processes; modeling in the time domain; forecasting; spectral analysis; linear systems. Appendix on Laplace, Fourier and Z-transforms; does not assume prior knowledge of transforms. New chapter on state-space models and the Kalman

filter. Other chapters revised to clarify and update. (*Third Edition*, TR, January 1986.) LB

Stochastic Processes, T(18: 2), P. *Time Series Analysis: Nonstationary and Noninvertible Distribution Theory.* Katsuto Tanaka. Ser. in Prob. & Stat. Wiley, 1996, x + 623 pp, $89.95. [ISBN 0-471-14191-7] Methods for noninvertible or nonstationary linear time series. Assumes background in mathematical statistics, including stationary stochastic processes. LB

Mathematical Statistics, T(16–17: 2), S, P*, L. *Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction.* Frank Lad. Ser. in Prob. & Stat. Wiley, 1996, xix + 484 pp, $69.95. [ISBN 0-471-14329-4] Somewhat mistitled. Mainly a formal development of material needed for mathematical statistics, primarily probabilistic, "as it is understood in the subjectivist perspective." This viewpoint, based in large part on the work of Bruno de Finetti, discounts much of current statistical practice on philosophical grounds. It is "set in the context of an operational and positivist approach to science, an analytic approach to philosophy, and a constructivist, finitist, and intuitionist approach to mathematics." Contains many historical and philosophical notes in making the case for this admittedly controversial minority view. RSK

Mathematical Computing, S(13–16). *Maple V Primer, Release 4.* Frank Garvan. CRC Pr, 1997, 143 pp, $14.95 (P). [ISBN 0-8493-2681-8] Pocket guide to Maple V Release 4. Lots of examples. MPR

Mathematical Computing, S*(14–15), C. *HP-48G/GX Investigations in Mathematics.* Donald R. LaTorre, Donald L. Kreider, T.G. Proctor. Charles River Media (POB 417, 403 VFW Dr., Rockland, MA 02370), 1996, xiii + 636 pp, $29.95 (P), with disk. [ISBN 1-886801-23-1] A detailed, rich, clearly written collection of guided investigations, using the HP-48, into areas of calculus, differential equations, engineering mathematics, and linear algebra. Includes many examples and (solved) exercises, sample programs, and a basic introduction to the machine. Disk contains a large collection of special-purpose programs transferable (with appropriate interface kit) to the HP-48. PZ

Computer Science, P. *Exploring Java.* Patrick Niemeyer, Joshua Peck. O'Reilly & Associates, 1996, xv + 407 pp, $24.95 (P). [ISBN 1-56592-184-4]

Computer Science, P. *Algebraic 3-D Modeling.* Andreas Hartwig. AK Peters, 1996, x + 222 pp, $59. [ISBN 1-56881-023-7] Mathematical treatment of geometric modeling, in particular boundary representations, in computer science. Aims for a generalized algebra based on Boolean set operations, hull constructions, and combinatorial techniques applicable to a wide range of modeling environments. Uses a small language modeler called GECO to compare different approaches. MPR

Applications (Fluid Mechanics), P. *Mathematical Problems in the Theory of Water Waves.* Eds: F. Dias, J.-M. Ghidaglia, J.-C. Saut. Contemp. Math., V. 200. AMS, 1996, xxiii + 235 pp, $55 (P). [ISBN 0-8218-0510-X] Proceedings of a 1995 workshop at CIRM in Luminy, France.

Applications (Physics), P. *Topics in Statistical and Theoretical Physics: F.A. Berezin Memorial Volume.* Eds: R.L. Dobrushin, *et al.* AMS Transl. Ser. 2, V. 177: Adv. in Math. Sci., V. 32. AMS, 1996, ix + 223 pp, $99. [ISBN 0-8218-0425-1] 12 papers by students and colleagues of Berezin.

Applications (Quantum Theory), T(18: 2), S, P. *Lie Groups, Lie Algebras, Cohomology and Some Applications in Physics.* José A. de Azcárraga, José M. Izquierdo. Mono. on Math. Physics. Cambridge Univ Pr, 1995, xvii + 455 pp, $100. [ISBN 0-521-46501-X] Largely self-contained, but assumes some knowledge of differential geometry, Cartan calculus, and quantum field theory. The authors are superb mathematicians and good writers. MU

Applications (Relativity), T(18: 2), S, P. *Global Lorentzian Geometry, Second Edition.* John K. Beem, Paul E. Ehrlich, Kevin L. Easley. Mono. & Textbooks in Pure & Appl. Math., V. 202. Marcel Dekker, 1996, xiv + 635 pp, $175. [ISBN 0-8247-9324-2] Contains additional material on stability, gravitational plane wave space-times, and the splitting problem. (*First Edition*, TR, May 1982.) MU

Applications, P. *Nonlinear Mathematics and Its Applications.* Ed: Philip J. Aston. Cambridge Univ Pr, 1996, vii + 256 pp, $24.95 (P). [ISBN 0-521-57676-8] 9 papers from a 1995 Spring School for postgraduate students at the University of Surrey. Applications in engineering, fluid dynamics, material science, and biology.

Reviewers

LB: Lynne Baur, Carleton; DB: David Bressoud, Macalester; LC: Laura Chihara, St. Olaf; JD: Jill Dietz, St. Olaf; PF: Paul Froeschl, Macalester; PG: Philip Gloor, St. Olaf; TH: Tom Halverson, Macalester; RSK: Richard S. Kleber, St. Olaf; LCL: Loren C. Larson, St. Olaf; RM: Richard Molnar, Macalester; MPR: Matthew P. Richey, St. Olaf; MU: Milton Ulmer, Carleton; PZ: Paul Zorn, St. Olaf.

THE AUTHORS

JUDITH V. GRABINER received her B.S. in Mathematics from Chicago and her Ph.D. in the History of Science from Harvard. She has written *The Origins of Cauchy's Rigorous Calculus* (MIT, 1981), *The Calculus as Algebra: J.-L. Lagrange, 1736–1813* (Garland, 1990), has won MAA awards for articles in the history of mathematics, and has been a Sigma Xi National Lecturer. She is Flora Sanborn Pitzer Professor of Mathematics at Pitzer College in Claremont, California, and especially enjoys sabbaticals in Scotland.

PAT TOUHEY did his graduate work in homotopy theory under Joe Roitberg at the C.U.N.Y. Graduate School. He resides in beautiful Northeast Pennsylvania with his wife, Maureen, and children, Kate and Pat. In addition to his administrative chores as department chair at College Misericordia, he enjoys shovelling snow, raking leaves, and mowing the lawn.

SCOTT MCCULLOUGH received his Ph.D. in 1987 at the University of California at San Diego under the beneficent direction of Jim Agler. He spent two years at Indiana University on his way to the University of Florida. He works in operator theory.

LEIBA RODMAN attended the Latvian State University in Riga, Latvia and Tel Aviv University in Israel, and received at the latter the Ph.D. degree in 1978 under the direction of Professor Israel Gohberg. He held teaching positions at Tel Aviv University, Arizona State University, University of California at San Diego (a brief one semester position), and the College of William and Mary (his current position). His research interests include operator theory, linear algebra, and related problems in systems and control.

CLARK KIMBERLING received his Ph.D. in 1970 from Illinois Institute of Technology, where the director of his dissertation was Abe Sklar. Aside from special points, lines, and curves in the plane of a triangle, Kimberling's interests include hundreds of sequences accessible at Neil Sloane's *Online Encyclopedia of Integer Sequences:* http://netlib.bell-labs.com/netlib/att/math/sloane/doc/eisol.html. Kimberling has written several biographical articles, including "Emmy Noether," this MONTHLY 79 (1972) 136–149. The Noether article tells how a large collection of letters penned by Richard Dedekind, Georg Cantor, and Georg Frobenius was sent to Kimberling—at a cost of $25.00! Kimberling kept the letters in a bank until 1995; now they repose in the archives of Technische Universität Braunschweig, where Dedekind was Professor of Mathematics.

MIKLÓS LACZKOVICH graduated from the Eötvös Loránd University (Budapest, Hungary) in 1971. He is now a full professor of mathematics at the same institution. His main research interest is real function theory. He received the 1993 Ostrowski prize for the solution of Tarski's "circle-squaring" problem.

MARCELO POLEZZI received his B.S. from the Instituto de Biociêcias, Letras e Ciências Exatas (Universidade Estadual Paulista) in 1995, with a First. Nowadays, he is preparing himself to be a M.Sc. in mathematics, and his research interests are in the theory of singularities of differentiable maps.

DETLEF LAUGWITZ received his doctoral degree at the University of Göttingen in 1954. He teaches at the University of Technology in Darmstadt, Germany. His main research interests have ranged from differential geometry and convexity to analysis and its history. Since the 1950's he has been engaged in the creation of a genetic approach to a modern theory of infinitesimals (or nonstandard analysis). He has written several textbooks in these fields. Birkhäuser has just (1996) published his book *Bernhard Riemann*, devoted to Riemann's mathematics, physics, and philosophy. An English translation of the book is in progress.

ANDRE TOOM started his mathematical career as a student in a mathematical circle and received several prizes at Olympiads. He spent more than twenty years at Moscow University, first as a student and then as a scientist. He received a prize for his research from the Moscow Mathematical Society. Six years ago he was allowed by the Soviet authorities to go abroad and went further than they expected. Now he is an associate professor at the University of the Incarnate Word and this keeps him busy as a teacher. Most of his research in the last few years is on random processes and their applications to statistical physics. His main hobby is criticism of mathematical education, which is well-known to members of some e-mail lists.

HARDY GRANT recently took early retirement from the Department of Mathematics and Statistics, York University, Toronto, and now lives in Ottawa, where he grew up. His degrees are from Queen's University (Kingston, Ont.) and McGill University (Montreal). His speciality at York was an undergraduate "Humanities" course on the history and cultural influence of mathematics, and his publications are in that area. His other passions include baseball, birding, bridge, recreational computer programming, and 78-r.p.m. records.

EDITOR'S ENDNOTES

Stimulated by Margaret Kleinfeld's article "Calculus: Reformed or Deformed?", MONTHLY 103 (1996) 230–232, James Sandefur wrote:

> Professor Kleinfeld, among other things, argued for teaching fewer applications in mathematics courses. She offered the analogy of a guide in a foreign country showing a cathedral. The group became bored and asked to see the shopping center, analogous to our students getting bored and asking for applications. I would like to take her analogy one step further. Having recently been to Europe and toured many cathedrals, I have had tour guides who only describe the cathedral, "This archway is 20 meters high, the stained glass window is... ." The tours I have enjoyed the most integrated the actual building with the history (when built, wars in which it was damaged, parts added, etc.), the art (the paintings on the ceiling were by so-and-so, with the style coming from...), and the purpose (this cathedral was built to house the...). Just as the cathedral is more than a collection of stone and stained glass, mathematics is more than a collection of formulas and abstract ideas. We should give our students an appreciation for the genius it took to come up with a particular new idea that may seem simple now, just as the perspective of a painting in a cathedral took genius at the time it was painted, although today that use of perspective is common. Our students should understand the historical context of certain mathematics, such as the applications that led to a new advancement, just as the guide should describe the historical development of the cathedral. But most importantly, what students will remember best is the mathematics they have had time to explore, just as I remember best the parts of the cathedral I explored on my own. So when the group asks to see the shopping center, it's not that they don't like the cathedral; it is that they don't like the guide.

The complete publication information for a recently-reviewed book (MONTHLY 103 (1996) 705) is: *In Search of Infinity*, by Naum Ya. Vilenkin. Translated by Abe Shenitzer with the editorial assistance of Hardy Grant and Stefan Mykytiuk. Birkhäuser, 1995, 145 pp., $24.50.

The Challenge to identify the cover illustration on the August 1995 issue (MONTHLY 102 (1995) 660) resulted in 49 replies to David Fowler. None of the respondents had seen the plot before; the following identified it correctly: Juan Arias-de-Reyna, Donald Bridges, Henry Edwards, David V. Feldman, Jöran Friberg, Dean P. Foster, Niels-Henrik Holstein-Rathlou, Peter Jones, Azzedine Kaced, Steve Kass, K. Robin McLean, John Mason, Joe Moser, Les Reid, Joel E. Rosenberg, Jeremy T. Tyson, and Hansklaus Rummler. Professor Fowler provided full details in his article "The Binomial Coefficient Function," MONTHLY 103 (1996) 1–17.

Victor Klee offered the following comments on James Angelos et al., "Packability of Five Spheres on a Sphere Implies Packability of Six," MONTHLY 103 (1996) 894–896:

> ...their result...goes back at least to K. Shütte and B. van der Waerden, *Math. Ann.* 123 (1951) 96–124. The relationship (of six points on the sphere to five points) that they establish has also been established for...12 points to 11 points, and there are conjectures in the work of Raphael Robinson concerning the same phenomenon for...the cases in which the number of points is 24, 48, 60, or 120. However, the conjecture for 24 has been disproved by Tarnai in Budapest.

Finally, the charming photo of the *Chez Math* sign in Maastricht in the December '96 issue (MONTHLY 103 (1996) 845) has a scrambled attribution. It was contributed by Evan Romer at Susquehanna Valley High School, Conklin, NY.

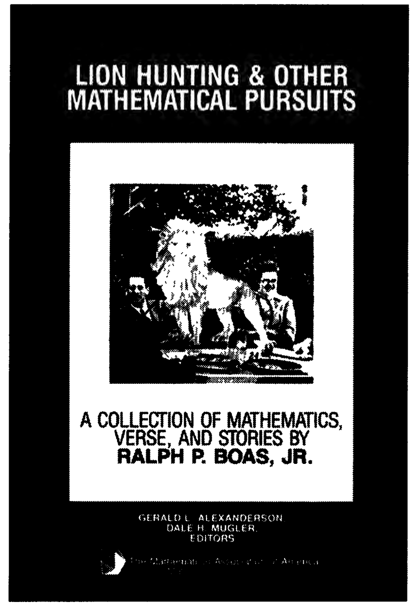
Roger A. Horn, *Editor*

Lion Hunting

and Other Mathematical Pursuits

A Collection of Mathematics, Verse, and Stories
by Ralph P. Boas, Jr.

Gerald L. Alexanderson and
Dale H. Mugler, Editors



I highly recommend Lion Hunting and Other Mathematical Pursuits to high school mathematics clubs, mathematics teachers of all levels, and anyone interested in mathematics. Perhaps the most important features of this book is how it subtly makes the reader aware of the nature of mathematics.

– The Mathematics Teacher

As a young man at the Institute for Advanced Study in Princeton, Ralph Philip Boas, Jr., together with a group of other mathematicians, published a light-hearted article on the "mathematics of lion hunting" under a pseudonym (1938). This sparked a sequence of articles on the topic, several of which are drawn together in this book.

Lion Hunting includes an assortment of articles that show the many facets of this remarkable mathematician, editor, writer, and teacher. Along with a variety of his lighter mathematical papers, the collection includes Boas' verse and short stories, many of which are appearing for the first time. Anecdotes and recollections of his numerous experiences and of his work and meetings with many distinguished mathematicians and scientists of his day are also included as well as photographs taken by Boas of Hardy, Littlewood, Besicovitch, Weil, and others.

The mathematical articles in this collection cover a range of topics. They include articles on infinite series, the mean value theorem, indeterminate forms, complex variables, inverse functions, extremal problems for polynomials and more. A special section of this book is devoted to articles about the teaching of mathematics, with titles such as "Calculus as an experimental science" and "Can we make mathematics intelligible?"

Boas's wit and playful humor are reflected in the verses included in this collection. The verses reflect the phases of his career as author, editor, teacher, department chair, and lover of literature. A section of the book describes the feud that Boas supposedly had with Bourbaki. Also included are many amusing anecdotes about famous mathematicians.

320 pp., Paperbound, 1995, ISBN 0-88385-323-X
List: $39.95 MAA Member: $28.95
Catalog Code: DOL-15,

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
P.O. Box 91112, Washington, DC 20090-1112
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ____________________

Address ____________________

City ____________________

State ____________ Zip ____________

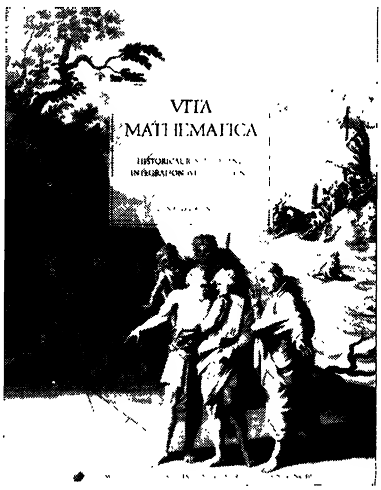
QTY.	CATALOG CODE	PRICE	AMOUNT
______	DOL-15	______	______
______	______	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________________ Expires ___/___

Signature ____________________

THE MATHEMATICAL ASSOCIATION OF AMERICA



Vita Mathematica

Historical Research and Integration with Teaching

Ronald Calinger, Editor

The use of the history of mathematics in the teaching of mathematics at all levels is an idea whose time has come. To use history in the teaching of undergraduate mathematics, the instructor must be familiar with the history as well as the mathematics. *Vita Mathematica* will enable college teachers to learn the relevant history of various topics in the undergraduate curriculum and help them incorporate this history in their teaching.

For example, should calculus be approached from a geometric or an algebraic point of view? The book shows us how two important eighteenth century mathematicians, Colin Maclaurin and Joseph-Louis Lagrange, understood the calculus from these different standpoints and how their legacy is still important in teaching calculus today. We also learn why Lagrange's algebraic approach dominated teaching in Germany in the nineteenth century. Some of the reasons for this are related to the appropriate foundations of the calculus, and so the book traces the ancient history of one of the possible foundations, the concept of indivisibles. Even though we generally do not use this concept formally today, many ideas for a heuristic approach to the calculus can be developed out of his study.

Vita Mathematica contains numerous other articles dealing with calculus, with algebra, combinatorics, graph theory, and geometry, as well as more general articles on teaching courses for prospective teachers. This volume, then, demonstrates that the history of mathematics is no longer tangential to the mathematics curriculum, but in fact deserves a central role.

Catalog Code: NTE40/JR
350 pp., Paperbound, 1996, ISBN 0-88385-097-4
List: $34.95 MAA Member: $29.00

Phone in Your Order Now! ☎ 1-800-331-1622

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders:** $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ____________________

Address ____________________

City ____________ State _____ Zip ________

Phone ____________________

Qty.	Catalog Code	Price	Amount
________	NTE40/JR	________	________

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling ________

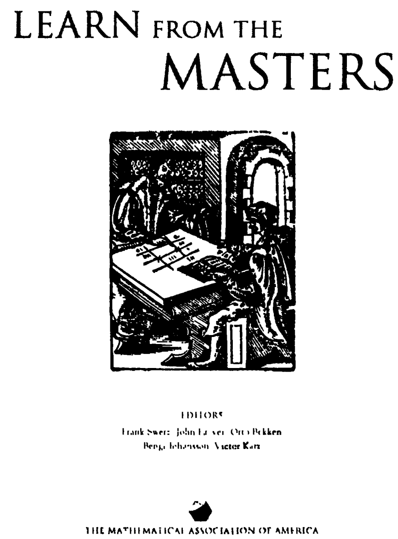
TOTAL ________

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________________ Expires ___/___

Signature ____________________

Learn from the Masters



Frank Swetz, John Fauvel, Otto Bekken, Bengt Johansson, Victor Katz, Editors

Provides high school and college teachers with important historical ideas and insights which can be immediately applied in the classroom.

This book is for college and high school teachers who want to know how they can use the history of mathematics as a pedagogical tool to help their students construct their own knowledge of mathematics. Often, a historical development of a particular topic is the best way to present a mathematical topic, but teachers may not have the time to do the research needed to present the material. This book provides its readers with historical ideas and insights which can be immediately applied in the classroom.

The book is divided into two sections: the first on the use of history in high school mathematics, and the second on its use in university mathematics. So, high school teachers planning a discussion of logarithms, will find here the historical background of that idea along with suggestions for incorporating that history in the development of the idea in class. College teachers of abstract algebra will benefit by reading the three articles in the book dealing with aspects of that subject and considering their ideas for presenting groups, rings, and fields.

The articles are diverse, covering fields such as trigonometry, mathematical modeling, calculus, linear algebra, vector analysis, and celestial mechanics. Also included are articles of a somewhat philosophical nature, which give general ideas on why history should be used in teaching and how it can be used in various special kinds of courses. Each article contains a bibliography to guide the reader to further reading on the subject.

This book grew out of a conference in Norway which brought together mathematicians and mathematics educators from a dozen countries who were interested in the use of the history of mathematics as a pedagogical tool in the teaching of mathematics. Since the conference which provided the genesis of this book took place in Norway near the home where Niels Henrik Abel spent his final days, the book's title comes from a note scribbled in one of Abel's notebooks: "It appears to me that if one wants to make progress in mathematics one should study the masters." The authors hope that readers will benefit from Abel's advice and show their students how they too can Learn from the Masters.

312 pp., Paperbound, 1995
ISBN-0-88385-703-0
List: $26.50
MAA Member: $20.95
Catalog Code: LRM/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ____________________

Address ____________________

City ____________________

State __________ Zip __________

QTY.	CATALOG CODE	PRICE	AMOUNT
____	LRM/JR	____	____
____	____	____	____
		TOTAL	____

Payment ☐ Check ☐ VISA ☐ MasterCard

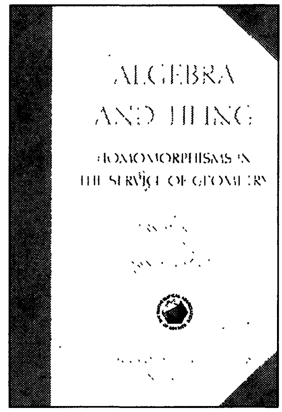
Credit Card No. ________________ Expires ___/___

Signature ____________________

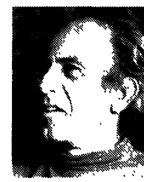
Algebra and Tiling

Homomorphisms in the Service of Geometry

Sherman Stein and Sándor Szabó



Sándor Szabó



Sherman Stein

Algebra and Tiling is perfect for bringing alive an abstract algebra course. Intuitive but difficult problems of geometry are translated into algebraic problems more amenable to solution. Full of nice surprises, the book is a pleasure to read.

—Choice

Often questions about tiling space or a polygon lead to other questions. For instance, tiling by cubes raises questions about finite abelian groups. Tiling by tripods or crosses raises questions about cyclic groups. From tiling a polygon with similar triangles, it is a short step to investigating automorphisms of real or complex fields. Tiling by triangles of equal areas soon involves Sperner's lemma from topology and valuations from algebra.

The first six chapters of *Algebra and Tiling* form a self-contained treatment of these topics, beginning with Minkowski's conjecture about lattice tiling of Euclidean space by unit cubes, and concluding with Laczkowicz's recent work on tiling by similar triangles. The concluding chapter presents a simplified version of Rédei's theorem on finite abelian groups: if such a group is factored as a direct product of subsets, each containing the identity element, and each of prime order, than at least one of them is a subgroup. A remarkable geometric implication of this result is developed in Chapter 2.

Algebra and Tiling is accessible to undergraduate mathematics majors, as most of the tools necessary to read the book are found in standard upper division algebra courses, but teachers, researchers and professional mathematicians will find the book equally appealing. Beginners will find the exercises and the material found in the appendices especially useful. The "Problems" section will appeal to both beginners and experts in the field. The book could serve as the basis of an undergraduate or graduate seminar or a source of applications to enrich an algebra or geometry course.

Contents

224 pp., Hardcover, 1994 ISBN 0-88385-028-1
List: $41.50 MAA Member $33.50
Catalog Code: CAM-25/JV

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name __________

Address __________

City __________

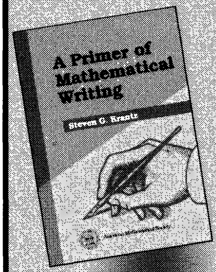
State __________ Zip __________

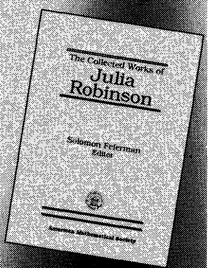
Qty.	Catalog Code	Price	Amount
____	CAM-25/JV	____	____
____	____	____	____
		TOTAL	____

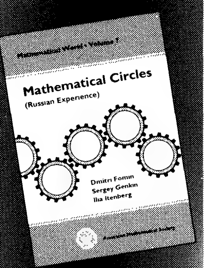
Payment ☐ Check ☐ VISA ☐ MasterCard

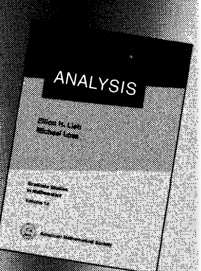
Credit Card No. __________ Expires ___/___

Signature __________









American Mathematical Society

Recently Published by the AMS

Analysis

Elliott H. Lieb, *Princeton University, NJ,* and **Michael Loss,** *Georgia Institute of Technology, Atlanta*

Not simply another book on real analysis, this straightforward, hands-on text provides readers at all levels—from beginning students to practicing analysts—with the basic concepts and standard tools necessary to understand analytical methods and better apply them to research in a variety of areas.

Analysis takes readers quickly from basic topics to applications (many of them quite deep), incorporating only those results and constructions that work successfully in mathematics and its applications. The authors take great care to include topics that any working analyst uses in everyday practice.

The book covers measure and integration, theory of L^pspaces, distribution theory, Fourier analysis, potential theory, Sobolev spaces, and much more. *Analysis* is a unique, practical book that everyone—from the graduate student, to the professional mathematician, to the physicist or engineer using analytical methods—will find interesting, stimulating, and useful.

Graduate Studies in Mathematics; Volume 14; 1996; 278 pages; Hardcover; ISBN 0-8218-0632-7; List $35; All AMS members $28; Order code GSM/14MM75

The Collected Works of Julia Robinson

Solomon Feferman, *Stanford University, CA,* Editor

This volume presents all the published works—spanning more than thirty years—of Julia Bowman Robinson. These papers constitute important contributions to the theory of effectively calculable functions and to its applications.

The volume also includes an extensive biographical memoir on the life and work of Robinson.

Collected Works; Volume 6; 1996; 338 pages; Hardcover; ISBN 0-8218-0575-4; List $69; Individual member $41; Order code CWORKS/6MM75

Lectures on the Mathematics of Finance

Ioannis Karatzas, *Columbia University, New York*

In this text, the author discusses the main aspects of mathematical finance. These include arbitrage, hedging and pricing of contingent claims, portfolio optimization, incomplete and/or constrained markets, equilibrium, and transaction costs. The book outlines advances made possible during the last fifteen years due to the methodologies of stochastic analysis and control. Readers are presented with current research, and open problems are suggested.

CRM Monograph Series; Volume 8; 1997; 148 pages; Hardcover; ISBN 0-8218-0637-8; List $39; Individual member $23; Order code CRMM/8MM75

Mathematical Circles

Dmitri Fomin, *St. Petersburg State University, Russia,* **Sergey Genkin,** *Microsoft Corporation,* and **Ilia Itenberg,** *Institut de Recherche Mathimatique de Rennes, France*

This is a sample of rich Russian mathematical culture written by professional mathematicians with great experience in working with high school students ... Problems are on very simple levels, but building to more complex and advanced work ... [contains] solutions to almost all problems; methodological notes for the teacher ... developed for a peculiarly Russian institution (the mathematical circle), but easily adapted to American teachers' needs, both inside and outside the classroom.

—***from the Translator's notes***

What kind of book is this? It is a book produced by a remarkable cultural circumstance in the former Soviet Union which fostered the creation of groups of students, teachers, and mathematicians called "mathematical circles". The work is predicated on the idea that studying mathematics can generate the same enthusiasm as playing a team sport—without necessarily being competitive.

Mathematical World; Volume 7; 1996; 272 pages; Softcover; ISBN 0-8218-0430-8; List $29; All AMS members $23; Order code MAWRLD/7MM75

A Primer of Mathematical Writing

Steven G. Krantz, *Washington University, St. Louis, MO*

This book is about writing in the professional mathematical environment. In many ways, this text complements Krantz's previous bestseller, *How to Teach Mathematics.* Those who are familiar with Krantz's writing will recognize his lively, inimitable style.

In this volume, he addresses these nuts-and-bolts issues:

- Syntax, grammar, structure, and style
- Mathematical exposition
- Use of the computer and TEX
- E-mail etiquette
- All aspects of publishing a journal article

Krantz's frank and straightforward approach makes this particularly suitable as a textbook. Readers will find in reading this text that Krantz has produced a quality work which makes evident the power and significance of writing in the mathematics profession.

1996; 223 pages; Softcover; ISBN 0-8218-0635-1; List $19; All AMS members $15; Order code PMWMM75

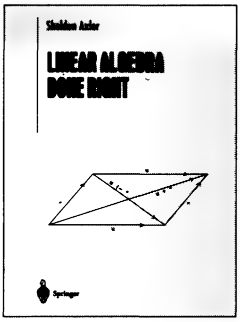
All prices subject to change. Charges for delivery are $3.00 per order. For air delivery outside of the continental U. S., please include $6.50 per item. *Prepayment required.* Order from: **American Mathematical Society,** P. O. Box 5904, Boston, MA 02206-5904. For credit card orders, fax (401) 331-3842 or call toll free 800-321-4AMS (4267) in the U. S. and Canada, (401) 455-4000 worldwide. Or place your order through the AMS bookstore at http://www.ams.org/bookstore/. Residents of Canada, please include 7% GST.

SPRINGER FOR MATHEMATICS

SHELDON AXLER, Michigan State University, East Lansing

LINEAR ALGEBRA DONE RIGHT

SECOND EDITION



"The most original linear algebra book to appear in years... a tour de force in the service of simplicity and clarity." – CHOICE

"The determinant-free proofsare elegant and intuitive." – AMERICAN MATHEMATICAL MONTHLY

This second edition of the highly-acclaimed text, *Linear Algebra Done Right*, has been extensively rewritten and updated with the addition of a number of illustrations and three new sections.

1997/APP. 251 PP, 22 ILLUS./SOFTCOVER/$29.00 (TENT.)
ISBN 0-387-98258-2
ALSO IN HARDCOVER: ISBN 0-387-98259-0/$49.00 (TENT.)
UNDERGRADUATE TEXTS IN MATHEMATICS

Instructor's Manual Available

MATHEMATICS FACULTY OF MOUNT HOLYOKE COLLEGE

LABORATORIES IN MATHEMATICAL EXPERIMENTATION

A Bridge to Higher Mathematics

The text is composed of a set of sixteen laboratory investigations which allow the student to explore rich and diverse ideas and concepts in mathematics. The approach is hands-on, experimental, an approach that is very much in the spirit of modern pedagogy. The course is typically offered in one semester, at the sophomore (second year) level of college. It requires completion of one year of calculus. The course provides a transition to the study of higher, abstract mathematics. Instructor's manual is also available (ISBN 0-387-94998-4, $12.95).

Contents: Iteration of Linear Functions • Iteration of Quadratic Functions • Iterated Linear Maps in the Plane • Iteration to Solve Equations • Cyclic Difference Sets • Prime Numbers • The Coloring of Graphs • The Euclidean Algorithm • The Euclidean Algorithm for Complex Integers • Randomzed Response Surveys • Parametric Curve Representation • Sequences and Series • p-adic Numbers • Some Periodic Functions • Numerical Integration and Integration by Chance • Polyhedra

1997/304 PP., 60 ILLUS./SOFTCOVER/$34.95
ISBN 0-387-94922-4
TEXTBOOKS IN MATHEMATICAL SCIENCES

H.-C. HEGE, Berlin, Germany and **K. POLTHIER**, Technische Universität Berlin, Germany (Eds.)

VISUALIZATION AND MATHEMATICS

Experiments, Simulations and Environments

Visualization methods rely heavily on mathematical concepts. Applications of visualization in mathematical research and the use of mathematical methods in visualization were topics of an international workshop in Berlin, June 1995. Based upon the workshop, selected contributions present topics of particular interest in current research. Experts report on the latest work, giving an overview of this fascinating new area and offering insight into state-of-the-art techniques for solving visualization problems and mathematical questions.

1997/APP. 384 PP./HARDCOVER/$99.00/ISBN 3-540-61269-6

New!

ALLAN J. ROSSMAN and **BARR VON OEHSEN**

WORKSHOP STATISTICS

Discovery with Data and the Graphing Calculator

Relying heavily on the use of graphing calculator technology *Workshop Statistics* uses real (and student-generated) data rather than canned examples and rewards conceptual understanding and clear communication over computational prowess. Data sets from the book (in TI-83 format) and TI-83 programs to accompany the text are available (ASCII, tab-delimited) via the internet at gopher://stats.dickinson.edu.

Over 90 classroom activities are grouped into six broad units:

- *Exploring Data:* Distributions includes measures of centers and spread
- *Exploring Data:* Relationships includes graphical displays of the correlation coefficient, least squares regression, and relationships with categorical variables
- *Exploring Data:* Randomness includes sampling distributions, confidence, significance, normal distributions and the Central Limit Theorem
- *Inference from Data:* Principles includes more about confidence intervals and tests of significance
- *Inference from Data:* Comparisons includes designing experiments and comparing proportions
- *Inference from Data:* Measurements includes inference for a population mean and comparing two means

1997/496 PP./SOFTCOVER/$29.95/ISBN 0-387-94997-6

Visit the Workshop Statistics website:
http://stats.dickinson.edu/math/rossman/wshome.html

New!

I.N. BRONSHTEIN and **K.A. SEMENDYAYEV**

HANDBOOK OF MATHEMATICS

This guidebook to mathematics contains in handbook form, the fundamental working knowledge of mathematics which is needed as an everyday guide for working scientists and engineers, as well as for students. Easy to understand, and convenient to use, this guide book gives concise information necessary to evaluate most problems which occur in concrete applications.

1997/988 PP., 95 ILLUS./HARDCOVER/$61.00 (TENT.)
ISBN 3-540-62130-X

For our latest textbooks please visit out textbook site at
http://www.springer-ny.com/math/text_books/

ORDER TODAY!

- **Call:** 1-800-SPRINGER or **Fax:** (201) 348-4505
- **Write:** Springer-Verlag New York, Inc., Dept. #S287, PO Box 2485, Secaucus, NJ 07096-2485
- **Visit:** Your local technical bookstore
- **E-mail:** orders@springer-ny.com
- **Instructors:** Call for info on textbook exam copies

5/97 Ref. #S287

Springer

http://www.springer-ny.com

THE AMERICAN MATHEMATICAL MONTHLY

Volume 104, Number 6 June–July 1997

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and Notes should be sent to the Editor:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send your email address and 3 copies of the complete manuscript (including all figures with captions and lettering), typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper.

Letters to the Editor on any topic are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

See the MONTHLY section of MAA Online for current information such as contents of issues, descriptive summaries of forthcoming articles, tips for authors, and preparation of manuscripts in TEX:

http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all problems/solutions material, typewritten on only one side of the paper.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Mr. JOE WATSON, Advertising Manager

Change of address, missing issues inquiries, and other subscription correspondence:
MAA Service Center
maahq@maa.org

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Approximate Isometries on Euclidean Spaces

Rajendra Bhatia and Peter Šemrl

1. INTRODUCTION. Let E and F be Banach spaces. An *isometry* from E to F is a map f: $E \to F$ such that

$$\|f(x) - f(y)\| = \|x - y\| \quad \text{for all } x, y \in E. \tag{1}$$

Every isometry is continuous and injective. Among the earliest theorems for Banach spaces is the Mazur-Ulam Theorem [**13**]. This says that if f is a surjective isometry between real Banach spaces E and F, and if $f(0) = 0$, then f is linear. The conclusion is not valid for complex Banach spaces (just consider the complex conjugation on $\mathbb{C}$). The hypothesis of surjectivity is essential in general, but can be dropped for a large class of Banach spaces that includes real Hilbert spaces. The condition $f(0) = 0$ is necessary for f to be linear. If f is any isometry then $f - f(0)$ is also an isometry, so this condition is no serious restriction.

If distances are known imprecisely one may not be able to say whether f is an isometry. Then the concept of an approximate isometry is useful. Given $\varepsilon > 0$, a map f: $E \to F$ is called an *ε-isometry* if

$$\big|\|f(x) - f(y)\| - \|x - y\|\big| < \varepsilon \quad \text{for all } x, y \in E. \tag{2}$$

Note that if f is an ε-isometry then so is $f - f(0)$. The following problem was posed by Hyers and Ulam [**9**]. If f is a surjective ε-isometry between real Banach spaces E and F such that $f(0) = 0$, then does there exist a surjective linear isometry g: $E \to F$ such that

$$\|f(x) - g(x)\| \le K\varepsilon \quad \text{for all } x \in E, \tag{3}$$

where the constant K is independent of f, but can depend on the spaces E and F? Hyers and Ulam [**9**] showed that if $E = F$ is a real Hilbert space then the answer is in the affirmative with $K \le 10$.

The Hyers-Ulam problem has been solved over the years. It is only recently that it was shown that the sharp value of K is 2 for all Banach spaces [**14**].

The aim of this note is to discuss some of these matters, to explain a part of the original Hyers-Ulam ideas, and to show how to extract some more results from them. One major issue of concern through the article is how essential the assumption of surjectivity of f is for the conclusions.

2. ISOMETRIES. Let E be any Banach space and let x, y be any two points of E. The *algebraic midpoint* of x and y is the vector $m(x, y) = (x + y)/2$. A *metric midpoint* of x and y is any point z of E that satisfies

$$\|z - x\| = \|z - y\| = \frac{1}{2}\|x - y\|. \tag{4}$$

The algebraic midpoint is always a metric midpoint. It is easy to see that if E is a Hilbert space there are no other metric midpoints for any pair of vectors x, y. This

is not always so in all Banach spaces. Here is an easy example:

Let E be the space $\mathbb{R}^2$ with the l_1-norm; i.e., if $x = (x_1, x_2)$ then $\|x\| = |x_1| + |x_2|$. Let $x = (1,0)$ and $y = (0,1)$. The algebraic midpoint of x and y is $(\frac{1}{2}, \frac{1}{2})$. This is at distance 1 from x and y. So are all points z of the form (t,t), where $0 \le t \le 1$. All these points are metric midpoints of x and y. A pictorial representation of this phenomenon might be helpful; see Figure 1. The unit ball of E is a diamond centered at the origin. Shift this diamond's center to $(1,0)$ and then to $(0,1)$. The intersection of the boundaries of these two diamonds is precisely the set of metric midpoints of x and y.

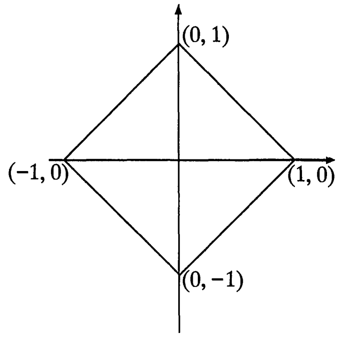


Figure 1

Let $M_0(x,y)$ be the set of all metric midpoints of x and y. It is easy to see that $M_0(x,y)$ is a closed, convex, and bounded subset of E.

There is a class of Banach spaces in which the norm is chosen so as to ensure that for all pairs x, y the set $M_0(x,y)$ is just the singleton $\{m(x,y)\}$. These are the *strictly convex* Banach spaces. The space E is called strictly convex, if whenever $\|x\| = \|y\| = 1$ and $\|(x+y)/2\| = 1$, then $x = y$ (that is, every point of the unit ball of E is an extreme point). For $1 < p < \infty$, l_p is strictly convex. A simple calculation with norms shows that if E is strictly convex then $M_0(x,y) = \{m(x,y)\}$ for all $x, y \in E$.

The importance of this observation is the following. The relation (4) that defines metric midpoints is unchanged under isometries, so if $f\colon E \to F$ is an isometry and if F is strictly convex then

$$f\left(\frac{x+y}{2}\right) = f(m(x,y)) = m(f(x), f(y)) = \frac{f(x)+f(y)}{2}.$$

Thus *every* isometry f from a Banach space E into a strictly convex Banach space F satisfies the equation

$$f\left(\frac{x+y}{2}\right) = \frac{f(x)+f(y)}{2}, \quad \text{for all } x, y \in E. \tag{5}$$

Now if $f(0) = 0$, this says that $f(x/2) = f(x)/2$ for all x. It follows, again from (5), that f is *additive*:

$$f(x+y) = f(x) + f(y), \quad \text{for all } x, y \in E. \tag{6}$$

It is clear from this equation that $f(nx) = nf(x)$, for every positive integer n. Also, choosing $y = -x$ in (6) we see that $f(x) = -f(x)$ for all x. Hence, $f(nx) = nf(x)$ for every integer n. Now it is easy to see that $f(rx) = rf(x)$ for every rational

number r. Since f is continuous, for all *real* α we have

$$f(\alpha x) = \alpha f(x) \quad \text{for all } x \in E.$$

Thus f is real linear even if the spaces E and F are complex; if they are real then f is linear. This proves the Mazur-Ulam Theorem in the special case when the space F is strictly convex. Note that in this case we did not require that f be surjective.

When the set $M_0(x, y)$ contains points other than $m(x, y)$, the preceding argument does not work. However, it is possible to give a metric characterization of the algebraic midpoint and then use a modified version of the above argument. Here is an outline of the argument.

Let x, y be any two points of a Banach space E. Starting with the set $M_0(x, y)$, define, inductively, for $n = 1, 2, \dots,$

$$M_n(x, y) = \left\{u \in M_{n-1}(x, y) \colon \|u - v\| \le \frac{d_{n-1}}{2} \quad \text{for all } v \in M_{n-1}(x, y)\right\}.$$

Here, $d_n = \operatorname{diam} M_n$. Then we have a nested sequence of closed sets $M_0(x, y) \supset M_1(x, y) \supset M_2(x, y) \supset \cdots$, with $\operatorname{diam} M_n \le d_0/2^n$. It is not difficult to prove that the point $m(x, y)$ is in $M_n(x, y)$ for all n. Hence,

$$\bigcap_{n=0}^{\infty} M_n(x, y) = \{m(x, y)\}. \tag{7}$$

This gives a metric characterization of the algebraic midpoint $m(x, y)$.

Now note that if f is a surjective isometry from a Banach space E onto a Banach space F, then

$$M_n(f(x), f(y)) = f(M_n(x, y)) \quad \text{for all } n.$$

At this step of the proof we do need to assume that f is surjective. If f were not surjective, we could have in F two points $f(x)$ and $f(y)$ whose metric midpoint is outside the range of f. So, from (7) we have

$$\left\{\frac{f(x) + f(y)}{2}\right\} = \bigcap_{n=0}^{\infty} M_n(f(x), f(y)) = \bigcap_{n=0}^{\infty} f(M_n(x, y)).$$

Since f is injective,

$$f\left(\bigcap_{n=0}^{\infty} M_n(x, y)\right) = \bigcap_{n=0}^{\infty} f(M_n(x, y)).$$

Now appealing to (7) again, we have

$$\frac{f(x) + f(y)}{2} = f\left(\frac{x + y}{2}\right).$$

As before, from this we can conclude that f is real linear. This proves the Mazur-Ulam Theorem.

Let us now give some simple examples to illustrate the necessity of the surjectivity assumption for general Banach spaces. Let $E = \mathbb{R}$ and let $F = \mathbb{R}^2$ with the ℓ_∞-norm; i.e., if $x = (x_1, x_2)$ then $\|x\| = \max(|x_1|, |x_2|)$. Let $f\colon E \to F$ be the map $f(t) = (t, \sin t)$. Since $|\sin t - \sin s| \le |t - s|$ for all t and s, it follows that f is isometric. Clearly f is not linear. To see another example, let $E = \mathbb{R}$ and let

$F = \mathbb{R}^2$ with the l_1-norm. Define the map $f\colon E \to F$ as

$$f(t) = \begin{cases} (t,0) & \text{if} \quad -1 \le t \le 1 \\ (-1, t+1) & \text{if} \quad t \le -1 \\ (1, t-1) & \text{if} \quad t \ge 1 \end{cases}.$$

This is the piecewise linear curve illustrated in Figure 2. It is easy to verify that

$$\|f(t) - f(s)\| = |t - s| \quad \text{for all } t, s \in \mathbb{R}.$$

Thus f is isometric, but not linear.

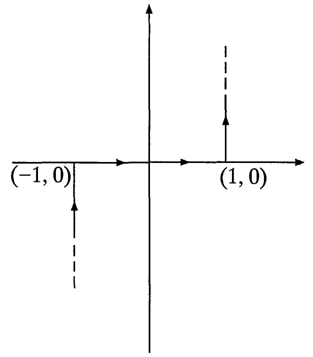


Figure 2

Can this phenomenon occur if $\dim E = \dim F$? The answer is no for finite-dimensional spaces. It was shown by Charzyński [**5**], [**6**, p. 143] that if E, F are n-dimensional real normed spaces then every isometry $f\colon E \to F$ satisfying $f(0) = 0$, is linear. Note that this implies that f is surjective.

Here is a simple proof of this theorem. Obviously, f maps S_E^r, the sphere of radius r centered at the origin of E into the sphere of the same type in F. Assume that there exists $r > 0$ such that $f(S_E^r)$ is a proper subset of S_F^r. Take any point $y \in S_F^r \setminus f(S_E^r)$. Then the restriction of f to S_E^r is an embedding of S_E^r into $S_F^r \setminus \{y\}$. If we have two different norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on $\mathbb{R}^n$ then every sphere with a positive radius r with respect to the norm $\|\cdot\|_1$ centered at 0 is homeomorphic to the unit sphere with respect to $\|\cdot\|_2$ (the homeomorphism can be defined by $x \mapsto x/\|x\|_2$ for every x with $\|x\|_1 = r$). Hence, the restriction of f to S_E^r can be considered as an embedding of the standard sphere S^{n-1} into the punctured sphere S^{n-1}, which is homeomorphic to $\mathbb{R}^{n-1}$. It is well-known that such embeddings do not exist. So, f must be surjective, and therefore, by the Mazur-Ulam theorem, it is linear. It is interesting to note that the name of Ulam is associated also with the theorem from topology used here. This is the Borsuk-Ulam Theorem; see [**12**, p. 170].

There is a more general version of the Mazur-Ulam Theorem that goes beyond Banach spaces to locally convex topological vector spaces. See [**6**, Chapter VII]. The idea of the proof is essentially the same, but now the algebraic midpoint is characterized in terms of prenorms.

3. APPROXIMATE ISOMETRIES. We have defined ε-isometries in Section 1 and explained the Hyers-Ulam problem. Since surjectivity of f is a necessary requirement in the Mazur-Ulam Theorem, it is natural to impose that condition here too. However, there is a significant difference between the two problems in

this respect. In Section 2 we explained how for a large class of Banach spaces (including Euclidean spaces) the Mazur-Ulam Theorem can be proved without this assumption. Hyers and Ulam gave an example of an ε-approximate isometry f from $\mathbb{R}$ into the Euclidean space $\mathbb{R}^2$, with $f(0) = 0$, that cannot be uniformly approximated by any linear isometry $g\colon \mathbb{R} \to \mathbb{R}^2$. They defined $f\colon \mathbb{R} \to \mathbb{R}^2$ by

$$f(t) = \begin{cases} (t, 0) & \text{if} \quad t \leq 1 \\ (t, c \log t) & \text{if} \quad t > 1 \end{cases}.$$

Then for each ε, we can choose c such that f is an ε-isometry. To see this note that $\log t$ is a concave function, and hence for $1 < s < t$,

$$\frac{\log t - \log s}{t - s} \leq \frac{\log t}{t - 1}.$$

Since $(\log t)/t \to 0$ as $t \to \infty$, this means that $\|f(t) - f(s)\|$ is asymptotically like $|t - s|$. More formally, it is an easy exercise to show that f is an ε-isometry whenever

$$\varepsilon > c^2 \max_{t>1} \left\{ \frac{(\log t)^2}{2t - 2} \right\}.$$

However, the set $\{\|f(t) - g(t)\| : t \in \mathbb{R}\}$ is unbounded for every linear isometry $g\colon \mathbb{R} \to \mathbb{R}^2$.

After the Hyers-Ulam solution of the problem for Hilbert spaces, there were several papers giving partial solutions for special Banach spaces. A breakthrough was made by Gruber [**8**], who proved that if a constant K satisfying (3) can be found (for a given pair of real Banach spaces E and F) then this inequality remains true if we choose $K = 5$. Further, he proved that this can always be done if E and F are finite-dimensional. In the general case of all real Banach spaces this was proved by Gevirtz [**7**]. Finally, it was shown by Omladič and Šemrl that the choice $K = 2$ works in (3) for all real Banach spaces E and F [**14**]. Here is a simple example that shows the inequality (3) with $K = 2$ is sharp. Define $f\colon \mathbb{R} \to \mathbb{R}$ by

$$f(t) = \begin{cases} t - 1 & \text{if} \quad t \notin [0, 1/2] \\ -3t & \text{if} \quad t \in [0, 1/2] \end{cases}.$$

One can easily check that f is a surjective 1-isometry satisfying $f(0) = 0$. The only linear isometries $g\colon \mathbb{R} \to \mathbb{R}$ are $g(t) = t$ and $g(t) = -t$. Obviously, the second one does not approximate f uniformly, while $\max|f(t) - t| = |f(\frac{1}{2}) - \frac{1}{2}| = 2$.

4. EUCLIDEAN SPACES. The Hyers-Ulam example explained in Section 3 can be modified to show that if E and F are real Hilbert spaces with either $\dim E < \dim F$, or $\dim E = \dim F = \infty$, then there exists an ε-isometry $f\colon E \to F, f(0) = 0$, that is not uniformly close to any linear isometry. Of course, such an f is not surjective.

What happens in the remaining case, $\dim E = \dim F < \infty$? The following theorem gives the answer.

Theorem 1. *Let E_n be an n-dimensional Euclidean space and let $f\colon E_n \to E_n$ be an ε-isometry satisfying $f(0) = 0$. Then there exists a unique bijective linear isometry*

$g\colon E_n \to E_n$ such that

$$\|f(x) - g(x)\| \le 2\varepsilon$$

for all $x \in E_n$.

Our proof has two steps. First we use two theorems from the Hyers-Ulam paper to find an isometry g and a constant K (depending on n) such that the inequality (3) is true. Then we use the special inner product structure of E_n to show that K can be replaced by 2. This argument is simpler than the one in **[14]** for arbitrary Banach spaces, and requires no assumption of surjectivity on f.

The inner product between two vectors x and y will be denoted by $\langle x, y\rangle$. Let $f\colon E_n \to E_n$ be an ε-isometry satisfying $f(0) = 0$. Assume for a moment that f can be uniformly approximated by a linear isometry $g\colon E_n \to E_n$, that is, there exists a positive real constant M such that

$$\|f(x) - g(x)\| \le M \quad \text{for all } x \in E_n.$$

Let m be an arbitrary positive integer. Replacing x in this inequality by $2^m x$, dividing the obtained inequality by 2^m, and using linearity of g we get

$$\left\| \frac{f(2^m x)}{2^m} - g(x) \right\| \le \frac{M}{2^m} \quad \text{for all } x \in E_n$$

and for all positive integers m. This shows that if

$$\lim_{m\to\infty} \frac{f(2^m x)}{2^m} \tag{8}$$

exists, then a linear isometry g can approximate f uniformly if and only if $g(x)$ is equal to this limit for every x. The sequence in (8) is now called the *Hyers-Ulam sequence*.

The first result in the Hyers-Ulam paper **[9]** states that this sequence does converge for every x.

Lemma 2. *Let E_n be an n-dimensional Euclidean space. Suppose that $\varepsilon > 0$ and that $f\colon E_n \to E_n$ is an ε-isometry satisfying $f(0) = 0$. Then*

$$g(x) = \lim_{m\to\infty} \frac{f(2^m x)}{2^m}$$

exists for every $x \in E_n$. The mapping g is a linear bijective isometry.

After this, Hyers and Ulam prove that an ε-isometry (not necessarily surjective) "approximately preserves" orthogonality, in the following sense:

Lemma 3. *Let f and g be as in Lemma 2, and let $u \in E_n$ be a unit vector. Then for every $x \in E_n$ orthogonal to u we have $|\langle f(x), g(u)\rangle| \le 3\varepsilon$.*

Proof of Theorem 1: Let $g\colon E_n \to E_n$ be as in Lemma 2. Since g^{-1} is an isometry, $g^{-1} \circ f$ is an ε-isometry. Note that $g^{-1} \circ f$ sends zero to zero and

$$\lim_{m\to\infty} \frac{(g^{-1} \circ f)(2^m x)}{2^m} = x$$

for all x. As it is enough to prove the conclusion for $g^{-1} \circ f$, we can assume with no loss of generality that $g(x) = x$ for every x.

First we show, using induction, the existence of a constant K (depending on n) such that $\|f(x) - x\| \le K\varepsilon$ for all x. Let f be an ε-isometry on E_1. Since $f(0) = 0$ we have $||f(x)| - |x|| < \varepsilon$ for all x. So, either $|f(x) - x| < \varepsilon$ or $|f(x) + x| < \varepsilon$.

For all x outside a large neighborhood of 0, only one of these can be true. It is now easy to find a constant K such that $|f(x) - x| \leq K\varepsilon$ for all x.

Assume now that we have already proved the assertion for $n - 1$ dimensional Euclidean spaces. Let x be any vector in E_n and let u be any unit vector orthogonal to x. By Lemma 3 with $g(y) \equiv y$ we have $|\langle f(x), u\rangle| \leq 3\varepsilon$. Let P be the orthoprojector onto $[u]^{\perp}$. For any $w \in [u]^{\perp}$ we define $f_1(w) = Pf(w)$. We claim that f_1 is a 7ε-isometry on $[u]^{\perp}$ satisfying $f_1(0) = 0$ and

$$\lim_{m\to\infty} \frac{f_1(2^m w)}{2^m} = w$$

for all w. Obviously, $f_1(0) = 0$. Next note that

$$\begin{aligned}
&\big|\|f_1(w) - f_1(w')\| - \|w - w'\|\big| \\
&\quad = \big|\|f(w) - \langle f(w), u\rangle u - f(w') + \langle f(w'), u\rangle u\| - \|w - w'\|\big| \\
&\quad \leq \big|\|f(w) - f(w')\| - \|w - w'\|\big| + 6\varepsilon \leq 7\varepsilon.
\end{aligned}$$

Finally,

$$\lim_{m\to\infty} \frac{f_1(2^m w)}{2^m} = \lim_{m\to\infty} \frac{Pf(2^m w)}{2^m} = Pw = w.$$

By the induction hypothesis, there exists a positive constant K_{n-1} such that

$$\|f_1(w) - w\| \leq 7K_{n-1}\varepsilon$$

for all $w \in [u]^{\perp}$. It follows that

$$\|f(x) - x\| = \|f_1(x) + \langle f(x), u\rangle u - x\| \leq 7K_{n-1}\varepsilon + 3\varepsilon.$$

Since x was an arbitrary vector, the induction step is over.

Now we will show how to replace K by 2. Take any $x \in E_n$ and set $\|f(x) - x\| = a$. Assume that $a \neq 0$. Denote by y the unit vector satisfying $f(x) - x = ay$. The vector x can be written as $x = x_0 + by$, $b \in \mathbb{R}$, where x_0 and y are orthogonal. For every positive integer m we have $f(x + my) = x + my + v_m$, where $\|v_m\| \leq K\varepsilon$ because of what we have shown in the first step. Write $v_m = b_m y + u_m$, $b_m \in \mathbb{R}$, where u_m and y are orthogonal. Consequently, $\|u_m\| \leq K\varepsilon$ and $|b_m| \leq K\varepsilon$. Using the fact that f is an ε-isometry with $f(0) = 0$ we have

$$\big|\|f(x + my)\| - \|x + my\|\big| < \varepsilon.$$

This can be rewritten as

$$\big|\|(m + b + b_m)y + (x_0 + u_m)\| - \|(m + b)y + x_0\|\big| < \varepsilon.$$

Since u_m is bounded and x_0 and u_m are orthogonal to y, this shows that for every $\mu > 0$ we have $|b_m| < \varepsilon + \mu$ if m is large enough.

Since f is an ε-isometry, we have

$$m - \varepsilon < \|f(x + my) - f(x)\| < m + \varepsilon,$$

or equivalently,

$$m - \varepsilon < \|(m - a + b_m)y + u_m\| < m + \varepsilon.$$

For large m this norm can be brought as close to $m - a + b_m$ as we wish. Since for large m we have $|b_m| < \varepsilon + \mu$ with μ being arbitrarily small, this is possible only if $a \leq 2\varepsilon$. ∎

We should remark that in the second part of our proof no reference was made to the finite dimensionality of the spaces involved. Thus, the factor 10 obtained by Hyers and Ulam (for the case of surjective isometries between infinite-dimensional Hilbert spaces) can be reduced to 2 using this argument.

It would be nice to extend Theorem 1 to ε-isometries $f\colon E \to F$ where E and F are arbitrary n-dimensional real normed spaces. In this case we have the following substitute for Lemma 2: There exists an increasing sequence (m_k) of positive integers such that

$$g(x) = \lim_{k\to\infty} \frac{f(m_k x)}{m_k} \tag{9}$$

exists for every $x \in E$. The mapping g is a linear bijective isometry. To prove this we first observe that the definition of ε-isometry implies that the sequence $(n^{-1}f(nx))$ is bounded for every $x \in E$. We choose a dense subset $\{z_1, z_2, \dots\}$ in E. Applying the Cantor diagonal procedure we can find an increasing sequence (m_k) of positive integers such that

$$g(z_p) = \lim_{k\to\infty} \frac{f(m_k z_p)}{m_k}$$

exists for every positive integer p. Using the definition of ε-isometry once again we see that (9) exists for every $x \in E$. Clearly, $g(0) = 0$. To prove that g is an isometry we replace x and y in (2) by $m_k x$ and $m_k y$, respectively. Dividing the obtained inequality by m_k and sending k to infinity we conclude that g is an isometry. We have already proved that g must be linear.

ACKNOWLEDGMENT. This work was supported by the Ministry of Science and Technology of Slovenia.

REFERENCES

1. D. G. Bourgin, Approximate isometries, *Bull. Amer. Math. Soc.* 52 (1946), 704–714.
2. D. G. Bourgin, Approximately isometric and multiplicative transformations on continuous function rings, *Duke Math. J.* 16 (1949), 385–397.
3. D. G. Bourgin, Two dimensional ε-isometries, *Trans. Amer. Math. Soc.* 244 (1978), 85–102.
4. R. D. Bourgin, Approximate isometries on finite dimensional Banach spaces, *Trans. Amer. Math. Soc.* 207 (1975), 309–328.
5. Z. Charzyński, Sur les transformations isométriques des espaces du type (F), *Studia Math.* 13 (1953), 94–121.
6. M. M. Day, *Normed Linear Spaces*, 3rd ed., Springer, Berlin 1973.
7. J. Gevirtz, Stability of isometries on Banach spaces, *Proc. Amer. Math. Soc.* 89 (1983), 633–636.
8. P. M. Gruber, Stability of isometries, *Trans. Amer. Math. Soc.* 245 (1978), 263–277.
9. D. H. Hyers and S. M. Ulam, On approximate isometries, *Bull. Amer. Math. Soc.* 51 (1945), 288–292.
10. D. H. Hyers and S. M. Ulam, On approximate isometries on the space of continuous functions, *Ann. of Math.* 48 (1947), 285–289.
11. J. Lindenstrauss and A. Szankowski, Non linear perturbations of isometries, *Asterisque* 131 (1985), 357–371.
12. W. S. Massey, *Algebraic Topology: An Introduction*, Springer, New York, 1977.
13. S. Mazur and S. Ulam, Sur les transformations isométriques d'espaces vectoriels normés, *C. R. Acad. Sci. Paris* 194 (1932), 946–948.
14. M. Omladič and P. Šemrl, On non linear perturbations of isometries, *Math. Ann.* 303 (1995), 617–628.

Indian Statistical Institute
New Delhi - 110016
India
rbh@isid.ernet.in

Faculty of Mechanical Engineering
University of Maribor
Smetanova 17
2000 Maribor, Slovenia
peter.semrl@uni-mb.si

How To Do MONTHLY Problems With Your Computer

István Nemes, Marko Petkovšek, Herbert S. Wilf, and Doron Zeilberger

Fortunately, on 20 *April* 1977, *all of this kludgery was rendered obsolete when I found a decision procedure for this problem.* (*A discrete analog to the Risch algorithm for indefinite integration.*)

—R. William Gosper, Jr., *Indefinite Hypergeometric Sums in MACSYMA* (1977)

1. INTRODUCTION. The problem of finding simple evaluations of major classes of sums that involve factorials, binomial coefficients, and their q-analogues, has been completely solved. Sums that have the rather general form specified in Section 3 can all be done algorithmically, that is to say, you can do them on your own PC. Your computer evaluates the sum as a simple formula, if that's possible, and gives you a proof that you can check, or gives you a proof that your sum cannot be "done" in simple closed form, if that is the case.

We first briefly describe the algorithms and the theory that have achieved this goal. Second, to illustrate both the scope of the method and the fact that in some interesting cases human intervention still helps, we show how these computer methods would have fared in attacking 27 problems that have appeared over the years in the Problems section of this MONTHLY.

It happens (coincidentally, of course) that three of the authors of this article (PWZ) have just written a book **[8]** that describes the theoretical foundations of the solution of this problem, and also gives the software by means of which everyone can perform these sums *sans peine* (almost).

2. THE METHODS. The methods that have achieved the complete solution of this class of problems are the following:

- Sister Celine's method **[1]**
- Gosper's algorithm **[3]**
- Zeilberger's algorithm `ct` ("creative telescoping") **[11]**
- Wilf and Zeilberger's WZ method **[9]**
- Petkovšek's algorithm `Hyper` **[6]**

Here is a brief description of the scope of each of these algorithms (full descriptions are in **[8]**). Computer programs, in *Maple* or *Mathematica* versions, that carry out each of these algorithms are available free at `http://www.cis.upenn.edu/~wilf/AeqB.html`.

Sister Celine's algorithm has been superseded by faster ones, but her work contains the original ideas on which the later algorithms have built. What it does can be stated pretty simply: it finds recurrences for hypergeometric summands. The fundamental theorem of this subject, which we state precisely in Section 3,

holds that every proper hypergeometric summand does indeed satisfy a recurrence relation. For instance, if, under your summation sign, there lurks

$$F(n,k) = \binom{n}{k}^2,$$

then her method informs you that

$$nF(n,k) - (2n-1)(F(n-1,k) + F(n-1,k-1)) \\ +(n-1)(F(n-2,k) - 2F(n-2,k-1) + F(n-2,k-2)) = 0.$$

Why did you want to know that? Well, if you sum this recurrence over all integer k, you'll find immediately (try it!) that the sum $f(n) = \sum_k \binom{n}{k}^2$ satisfies $f(n) = 2(2n-1)f(n-1)/n$, and so by induction, $f(n) = \binom{2n}{n}$, and you have evaluated your sum. But, you say, you already knew that the sum of the squares of the binomial coefficients is $\binom{2n}{n}$? Sure you did, but the same method works on any sum of factorials and binomial coefficients and powers in the world, provided it's of the form described in Section 3. So it wasn't finding that one particular sum that was the revolutionary event. It was the fact that Sister Celine's method can find recurrences satisfied by any one of a huge class of summands, and, as was realized much later, from the recurrence for the summand there comes the recurrence for the sum, and from that comes the closed form evaluation of the sum, if it has one. We now have algorithms that handle all of those pieces.

Gosper's algorithm completely solves the problem of *indefinite* hypergeometric summation. Given a summand $F(k)$ that is a *hypergeometric term* in k (i.e., $F(k+1)/F(k)$ is a rational function of k), Gosper's algorithm finds a hypergeometric term $G(k)$ such that $F(k) = G(k+1) - G(k)$, if one exists, or prove that none exists, if that be the case. Thus it solves the discrete analogue of the antidifferentiation problem: instead of exhibiting a given *integrand* as the *derivative* of something, thereby enabling *integration* in finite terms, it exhibits a given *summand* as the *difference* of something, thereby enabling *summation* in finite terms. Examples of the operation of this algorithm are in Section 5.

Zeilberger's algorithm `ct` finds a recurrence for a given hypergeometric summand $F(n,k)$. To that extent, it solves the same problem that Sister Celine's method solves. The form of the recurrence that it finds is different, however, and that allows an enormous speedup in its operation time. His algorithm finds a recurrence for $F(n,k)$ in the form

$$\sum_{j=0}^{d} a_j(n)F(n+j,k) = G(n,k+1) - G(n,k), \tag{1}$$

in which G/F is a rational function (which the output exhibits) and the $a_j(n)$'s are polynomials in n. The power of this result derives from the fact that if we sum both sides of this recurrence over a certain range of k, the sum on the right side telescopes, and so is easy to handle, and we obtain a recurrence for the sum, $\sum_k F(n,k)$, that we are trying to deal with. The fundamental theorem guarantees that such recurrences always exists if F is a proper hypergeometric summand (see Section 3).

Wilf and Zeilberger's WZ method is at once a special case and a generalization of Zeilberger's method. In order to prove an identity of the type $\sum_k F(n,k) = 1$, it finds a recurrence of the form

$$F(n+1,k) - F(n,k) = G(n,k+1) - G(n,k), \tag{2}$$

where G/F is a rational function called the *proof certificate* of the identity. This form is clearly a special case of (1) above. A recurrence of this form does not always exist. When it does, one gets two benefits: first a very short proof of one's summation identity, and second, because of the symmetry of (2) in F and G, one finds a new identity, involving G, from the original one, involving F.

Petkovšek's algorithm `Hyper` finds closed form solutions $f(n)$ to linear difference equations with polynomial coefficients,

$$\sum_{j=0}^{d} a_j(n) f(n+j) = 0,$$

when such solutions exist, or it proves that they do not exist, when they do not. We use the phrase "closed form" in the following precise sense: $f(n)$ is said to be of (hypergeometric) *closed form* if it is equal to a linear combination of a fixed number, r, say, of hypergeometric terms in n. Thus `Hyper` completes the job of doing the summation problem because the methods just described, while they are guaranteed to give you a recurrence for your unknown sum, are not guaranteed to give you one of minimum order! But `Hyper` knows how to solve such recurrences in closed form, or to prove the impossibility of solving the recurrence in closed form, if that be the case.

3. THE THEORY. Are these algorithms just more tricks, that might or might not work, to try on sums? Quite the contrary. In fact, the algorithms are accompanied by theorems that precisely describe circumstances under which they are *guaranteed* to work. So these are definitely not of the let's-see-if-it-works genre. They *will* work if the hypotheses of the relevant theorems are satisfied.

We are talking about sums of the form $f(n) = \sum_{k=a(n)}^{b(n)} F(n,k)$. The whole method rests on the fact that if $F(n,k)$ is a suitable summand then it satisfies a recurrence relation of a certain form. A summand $F(n,k)$ is suitable (*proper hypergeometric*) if it is of the form

$$F(n,k) = P(n,k) \frac{\prod_{i=1}^{I} (a_i n + b_i k + c_i)!}{\prod_{j=1}^{J} (u_j n + v_j k + w_j)!} x^n y^k, \tag{3}$$

in which

- $P(n,k)$ is a polynomial in n and k, whose degree is a specific integer, and
- the limits I, J on the products are fixed specific nonnegative integers, and
- the quantities a_i, b_i, u_j, v_j are specific integers, and
- the quantities c_i, w_j, x, y may depend on parameters.

Suppose we have a summand of that kind. What can we expect?

Theorem 1. *If $F(n,k)$ is proper hypergeometric then there exist a nonnegative integer d, a rational function $R(n,k)$, and polynomials $\{p_j(n)\}_{j=0}^{d}$, independent of k, such that $F(n,k)$ satisfies*

$$\sum_{j=0}^{d} p_j(n) F(n+j,k) = G(n,k+1) - G(n,k)$$

where $G(n,k) = R(n,k)F(n,k)$.

This theorem of Zeilberger, and the `creative telescoping` algorithm that carries it out, are used to find recurrences for given sums. The existence part of

the proof follows from an earlier algorithm of Sister Mary Celine Fasenmyer that was used by her to find recurrences for hypergeometric polynomials.

From the recurrence for the summand one gets a recurrence for the sum. From the recurrence for the sum one gets the evaluation of the sum in closed form, if possible, or a proof of impossibility. The latter follow from algorithm `Hyper`, if the recurrence obtained is of order greater than 1. Just as the problem of finding recurrences has a life of its own, aside from its uses in evaluating sums, so algorithm `Hyper` has a life aside from finding out if sums have closed forms. Combinatorics is full of enumeration problems that lead to recurrences. With `Hyper` we can now find solutions of these, or else prove that closed forms do not exist, for the first time. In this way a large number of combinatorial sequences have been proved not to be of closed form, such as those in the following theorem.

Theorem 2. *None of the following famous sequences can be expressed in hypergeometric closed form*:

- *the sum of the cubes (also the fourth and fifth powers) of the binomial coefficients of order n,*
- *the number of $3 \times n$ Latin rectangles,*
- *the number of involutions on n letters,*
- *the derangement numbers,*
- *the sum of the first n of the binomial coefficients of order pn ($p > 2$)* [**7**],

For the whole story of this remarkable current of mathematical thought, see [**8**].

4. THREE RECIPES FOR SUCCESS. Given a sum $S(n) = \sum_{k=a(n)}^{b(n)} F(n, k)$ with a proper hypergeometric summand, Zeilberger's algorithm `ct` yields a linear recurrence relation $\mathscr{L}$ with polynomial coefficients, of order $d \geq 0$, satisfied by $S(n)$. This is very helpful in the following situations that interest us here (and in many other situations too):

1. To prove that $S(n) = t_n$ where t_n is given in closed form, simply verify that t_n also satisfies $\mathscr{L}$, and that it agrees with $S(n)$ for d sufficiently large consecutive values of n.
2. To prove equality of two such sums use algorithm `ct` on both, and find a common multiple, $\mathscr{M}$, of the two resulting recurrences. If the order of $\mathscr{M}$ is m, verify that the two sums agree for m sufficiently large consecutive values of n.
3. To find a closed form evaluation of $S(n)$, note first that if $d = 0$, or $d = 1$ and $\mathscr{L}$ is homogeneous, then such an evaluation is immediate from $\mathscr{L}$. Otherwise, for various special reasons we might be able to solve $\mathscr{L}$ by inspection; it might be homogeneous with constant coefficients, for instance. But if no solutions are immediately apparent, then
   (a) If $d = 1$ and $\mathscr{L}$ is inhomogeneous, then $S(n)$ can be expressed in terms of an "indefinite" sum which Gosper's algorithm will put into closed form provided such a form exists.
   (b) Otherwise, use `Hyper` to find all closed form solutions of $\mathscr{L}$. Homogenize first if $\mathscr{L}$ is inhomogeneous. If you are lucky and $\mathscr{L}$ is satisfied by $t > 0$ linearly independent hypergeometric terms, then:
      i. If $t = d$, any solution of $\mathscr{L}$ can be put into closed form by choosing an appropriate linear combination of hypergeometric solutions.

ii. If $t < d$, try to find a linear combination of hypergeometric solutions that agrees with $S(n)$ for d sufficiently large consecutive values of n.

iii. If this fails, use hypergeometric solutions to reduce the order of $\mathscr{L}$. Repeat these steps with the new recurrence.

This procedure is *guaranteed* to decide whether $S(n)$ has a closed form evaluation (and to find it when it exists) whenever $F(n, k)$ is proper hypergeometric, and the limits of summation are either infinite (recall that summands often have compact support) or linear in n. But sometimes, with a little help, it works even when $F(n, k)$ is not proper hypergeometric (cf. problems E 3258, 10206, 10223, 10388 in the next Section). We refer the interested reader to [**8**, Chapter 8] for more details.

During the reviewing of this paper, one of the readers asked us for an example of a problem that we had not been able to do by these methods, even though it may have appeared to be a candidate. Of course such a problem would have to violate the hypotheses of Theorem 1, while at the same time seeming, at first glance anyway, to satisfy them. A good example of such a problem is an identity whose truth was conjectured by Borwein and Bradley, and which has recently been proved by Almkvist and Granville. It states that

$$\frac{5}{2}\sum_{k=1}^{n}\binom{2k}{k}\frac{k^2}{4n^4+k^4}\prod_{j=1}^{k-1}\frac{n^4-j^4}{4n^4+j^4}=\frac{1}{n^2}, \qquad (n \geq 1).$$

After factoring the fourth degree polynomials that appear in the summand one discovers that it has exactly the form (3), except that, for instance, one of the numbers a_i is $\sqrt{-1}$, which is not a specific integer, so the conditions are not satisfied.

5. PROBLEMS AND SOLUTIONS. We looked through MONTHLY problems on sums and recurrences that have been published since 1978, and selected 27 of the kind we're considering here. We used algorithms `ct` and `Hyper`, following the recipes given in the previous section. Wherever possible we used Gosper's algorithm and the WZ method, which technically are special cases of Zeilberger's algorithm `ct` corresponding to $d = 0$, and to $d = 1$ with given closed form evaluation, respectively (d being the order of the resulting recurrence). Besides our own implementations, we used the outstanding implementation of Zeilberger's algorithm in *Mathematica* by P. Paule and M. Schorn [**5**], which excels especially when the resulting recurrence is not homogeneous.

Many of the problems were solved completely automatically, while others required a little human help. For example, in several sums that involve the floor function we humans carried out the replacement

$$\sum_k F(n,k,\lfloor k/2\rfloor) = \sum_k \left(F(n,2k,k) + F(n,2k+1,k)\right) \tag{4}$$

in which, if $F(n, k, m)$ is hypergeometric in n, k, m, the summand on the left *is not* hypergeometric, but the one on the right *is*. Other examples of human intervention include the choice of the "best" recurrence variable when the summand depends on more than one parameter, etc.

A notable exception in the amount of necessary human aid is the double sum in Problem E 3376, which required the sharp eyes of P. Paule [**4**] to notice a special relationship among the coefficients of the recurrence. In principle, of course,

multiple sums can be handled by the methods of [10], but it is nice to be able to "do" a double sum with single-sum methods.

And now, here are the problems and their solutions!

Problem 6407. (*Proposed in* 1982, *p.* 703; *solution in* 1984, *p.* 315)

Define $\begin{bmatrix} n \\ k \end{bmatrix}$ *by means of the relation*

$$\begin{bmatrix} n \\ k \end{bmatrix} = F_{n,k}/F_{k,k} \quad F_{n,k} = (q^n - 1)(q^{n-1} - 1)\cdots(q^{n-k+1} - 1),$$

so that $\begin{bmatrix} n \\ k \end{bmatrix}$ *is the so-called Gaussian polynomial. Prove the identity*

$$\sum_{k=1}^{n} \frac{q^k}{1 - q^k} = \sum_{k=1}^{n} \frac{(-1)^{k-1}}{1 - q^k} q^{k(k+1)/2} \begin{bmatrix} n \\ k \end{bmatrix}. \tag{5}$$

Denote the right side of (5) by $S(n)$. The q-version of algorithm `ct` yields the recurrence $S(n) - S(n-1) = q^n/(1 - q^n)$, which is also satisfied by the left side of (5). As they agree for $n = 0$, the identity is proved.

Problem E 3021. (*Proposed in* 1983, *p.* 645; *solution in* 1986, *p.* 652)

Let

$$p_n(x) = \sum_{k=0}^{n} \binom{n}{k}^2 (1 + x)^k (1 - x)^{n-k}. \tag{6}$$

Express $p_n(x)$ *as an explicit function of* $1 - x^2$.

We provide a partial solution as follows. By algorithm `ct`

$$4(n + 1)x^2 p_n(x) - 2(2n + 3)p_{n+1}(x) + (n + 2)p_{n+2}(x) = 0. \tag{7}$$

With $p_0(x) = 1$ and $p_1(x) = 2$, we see from this recurrence that $p_n(x)$ is a polynomial in x^2, and hence in $1 - x^2$. By comparing (7) with the three-term recurrence

$$(n + 1)P_n(x) - (2n + 3)xP_{n+1}(x) + (n + 2)P_{n+2}(x) = 0$$

satisfied by the Legendre polynomials $P_n(x)$, we find that $p_n(x) = (2x)^n P_n(1/x)$.

Problem E 3022. (*Proposed in* 1983, *p.* 645; *solution in* 1986, *p.* 736)

Show that, for any $\alpha > 0$ *and any positive integer* N,

$$\sum_{k=1}^{N} (-1)^{k-1} \binom{N}{k} \frac{k}{1 + (k-1)\alpha} = \prod_{k=1}^{N-1} \left(\frac{k+1}{k + \alpha^{-1}} \right).$$

We are to show that $\sum_k F(N, k) = 1$, where

$$F = \frac{(-1)^{k-1}(\alpha^{-1})^{\overline{N}}}{(k-1)!(N-k)!(k - 1 + \alpha^{-1})},$$

and $(x)^{\overline{n}}$ is the rising factorial. The WZ method does this with the rational proof certificate $(k - 1)(\alpha^{-1} + k - 1)/(N(k - N - 1))$, and a check of the case $N = 1$.

Problem E 3065. (*Proposed in* 1984, *p.* 649; *solution in* 1987, *p.* 378)

Let $n \geq 0$ *be any integer and let* k *be any integer such that* $k \geq n + 1$. *Then find a closed formula for*

$$\sum_{j=0}^{n} \frac{(-1)^j}{j+1} \binom{k}{j} \binom{k - 1 - j}{n - j}.$$

Let $S(n)$ be the sum in question. Algorithm ct finds that

$$S(n) + S(n+1) = \frac{\binom{k}{n+1}}{n+2},$$

whence

$$S(n) = (-1)^n \left(\sum_{j=1}^{n} \frac{(-1)^j}{j+1} \binom{k}{j} + S(0) \right) = (-1)^n \sum_{j=0}^{n} \frac{(-1)^j}{j+1} \binom{k}{j}.$$

Gosper's algorithm gives the final answer

$$S(n) = \frac{1}{k+1} \left(\binom{k}{n+1} + (-1)^n \right).$$

Problem E 3088. (*Proposed in* 1985, *p.* 359; *solution in* 1987, *p.* 685)

Show that, for every positive integer n,

$$\sum_{k=1}^{n} \frac{k \cdot k!}{n^k} \binom{n}{k} = n. \tag{8}$$

Let t_k denote the summand in (8). Gosper's algorithm finds that $t_k = s_{k+1} - s_k$ where $s_k = -nt_k/k$. Summing this recurrence on k from 1 to n gives the sum as $s_{n+1} - s_1 = n$.

Problem 6519. (*Proposed in* 1986, *p.* 403; *solution in* 1988, *p.* 156)

Let

$$F(a,b,m,n) = \sum_{k=0}^{m} \binom{a+m+n-2k}{n-k} \binom{a+n}{k} \binom{b+m}{m-k}, \tag{9}$$

where m and n are nonnegative integers. Show that $F(a,b,m,n) = F(b,a,n,m)$.

For $F(a,b,m,n)$ and $F(b,a,n,m)$ we compute recurrence relations with respect to n using algorithm ct. As it turns out, both sums satisfy the same recurrence of order 4:

$$\begin{aligned}
&-3(1+a+n)(2+a+n)(3+a+n)S(n)\\
&+2(2+a+n)(3+a+n)(10+a+b-2m+4n)S(n+1)\\
&(3+a+n)(53+5a+5b-ab-26m-4am-4bm\\
&\qquad\qquad +36n+2an+2bn-8mn+6n^2)S(n+2)\\
&-(7+a+b+2n)(1+4a+4b+ab+8m+am\\
&\qquad\qquad +bm+an+bn+2mn)S(n+3)\\
&+(4+n)(4+b+n)(4+a+b+n)S(n+4) = 0.
\end{aligned}$$

Checking that $F(a,b,m,n) = F(b,a,n,m)$ for $0 \le n \le 3$ therefore completes the proof.

Problem E 3190. (*Proposed in* 1987, *p.* 181: *solution in* 1988, *p.* 877)

Show that

$$\sum_{r=0}^{j} \frac{(-1)^r (N-2r)\binom{j}{r}}{(N-r)\cdots(N-r-j)} = 0 \tag{10}$$

for $j > 0$ *and* $N > 2j$.

Let t_r denote the summand in (10). Gosper's algorithm finds that $t_r = s_{r+1} - s_r$ where $s_r = r(r+j-N)t_r/(j(N-2r))$. Summing this recurrence on r from 0 to $j-1$ gives the sum as $s_j - s_0 + t_j = 0$.

Problem E 3207. (*Proposed in* 1987, *p.* 456; *solution in* 1990, *p.* 67)

If m is a positive integer, let

$$F_m(x) = \sum_{k=0}^{m} \left(1 - \frac{k}{m}\right)\binom{2m}{k} x^k (1-x)^{2m-k}.$$

Show that

$$F_{m-1}(x) - F_m(x) = \frac{1}{2(2m-1)}\binom{2m}{m} x^m (1-x)^m \quad \text{for } m > 1. \tag{11}$$

Zeilberger's algorithm `ct` instantly yields (11).

Problem E 3258. (*Proposed in* 1988, *p.* 259; *solution in* 1989, *p.* 651)

Prove that

$$\sum_{j=0}^{n} \binom{n}{j} 2^{n-j} \binom{j}{\lfloor j/2 \rfloor} = (2n+1 \quad n).$$

If we use the transformation (4) here, the sum in question becomes

$$\sum_j \frac{n+2}{2j+1}\binom{n}{2j} 2^{n-2j-1}\binom{2j+1}{j}.$$

If we divide the summand by the claimed right side, $\binom{2n+1}{n}$, the WZ method proves the identity with a proof certificate of $4j(j+1)/((2n+3)(2j-n-1))$.

Problem E 3335. (*Proposed in* 1989, *p.* 525; *solution in* 1990, *p.* 927)

Solve the recurrence

$$x_0 = a,\ x_1 = b,\ x_{n+2} = x_{n+1} + x_n/(n+1) \quad \text{for } n = 0, 1, 2, \ldots$$

both exactly (*in terms of familiar functions of n*) *and asymptotically.*

Algorithm `Hyper` gives one solution, $n+1$. Then by reducing the order we find that $(n+1)\sum_{k=0}^{n}(-1)^k/(k+1)!$ is another. So

$$x_n = (n+1)\left(a + (b-2a)\sum_{k=0}^{n+1}\frac{(-1)^k}{k!}\right),$$

which is asymptotic to $(n+1)(a + (b-2a)/e)$.

Problem E 3352. (*Proposed in* 1989, *p.* 838; *solution in* 1991, *p.* 369)

Show that

$$\sum_{n=0}^{\infty} \frac{1}{n!(n^4+n^2+1)} = \frac{e}{2}.$$

This is equivalent to

$$\sum_{n=0}^{\infty}\left(\frac{1}{n!(n^4+n^2+1)} - \frac{1}{2n!}\right) = 0. \tag{12}$$

Let t_n denote the summand in (12). Gosper's algorithm finds that $t_n = s_{n+1} - s_n$ where $s_n = n^2/(2n!(n^2-n+1))$. Summing this recurrence on n from 0 to ∞ gives the sum as $s_\infty - s_0 = 0$.

Problem E 3376. (*Proposed in* 1990, *p.* 240; *solution in* 1992, *p.* 63)

Prove that

$$\sum_{i=0}^{N}\sum_{j=0}^{N}\binom{i+j}{j}^2\binom{4N-2i-2j}{2N-2j} = (2N+1)\binom{2N}{N}^2,$$

for any positive integer N.

Let $S(N, i)$ denote the inner sum on the left. Algorithm ct finds the recurrence

$$\sum_{j=0}^{3} p_j(N,i)S(N,i+j) = 0 \tag{13}$$

where

$$\begin{aligned}
p_0(N,i) &= (1+i)(-i+N)(-1-2i+2N),\\
p_1(N,i) &= -18-32i-22i^2-6i^3-11N+4iN+8i^2N-30N^2-20iN^2,\\
p_2(N,i) &= (2+i)(27+23i+6i^2+9N-4iN+18N^2),\\
p_3(N,i) &= -2(2+i)(3+i)^2.
\end{aligned}$$

Following Paule [4], we notice that $\Sigma_{j=0}^{3} p_j(N, i-j) = -2(2N+1)^2$ is independent of i. Summing recurrence (13) on i from -3 to N, and changing the order of summation to take advantage of this, we obtain

$$\sum_{i=1}^{3}\sum_{j=0}^{3-i}\big(S(N,-i)p_j(N-i-j) + S(N,N+i)p_{j+i}(N,N-j)\big)$$
$$-2(2N+1)^2\sum_{i=0}^{N} S(N,i) = 0. \tag{14}$$

Since for positive integer i the sums $S(N-i)$ and $S(N, N+i)$ contain only i nonzero terms, the result

$$\sum_{i=0}^{N} S(N,i) = (2N+1)\binom{2N}{N}^2$$

can be readily computed from (14).

Problem E 3439. (*Proposed in* 1991, *p.* 437; *solution in* 1993, *p.* 188)

If M and N are nonnegative integers, prove that

$$\binom{M+N}{M} = \sum_{0\le a\le \frac{M-1}{2}}\binom{M-a-1}{a}\binom{N+a}{2a+1} + \sum_{0\le a\le \frac{M}{2}}\binom{M-a}{a}\binom{N+a}{2a}. \tag{15}$$

If $M = 0$ both sides are 1. When $M > 0$ the two sums on the right can be combined into a single hypergeometric sum

$$S(N) = \sum_{a=0}^{M-1}\frac{MN-2aN+aM+M-a}{(2a+1)(M-a)}\binom{M-a}{a}\binom{N+a}{2a}$$

for which creative telescoping finds the recurrence

$$(N+1)S(N+1) - (M+N+1)S(N) = 0$$

satisfied by $\binom{M+N}{M}$. As the two sides of (15) agree at $N = 0$, the identity is proved.

Problem 10206. (*Proposed in* 1992, *p.* 266; *solution in* 1995, *p.* 657)

If m and k are positive integers, prove that

$$\sum_r\binom{r}{k-r}\binom{m}{r} = \sum_j\binom{\lfloor j/2\rfloor}{k-j}\binom{m-k+\lfloor 3j/2\rfloor}{j}. \tag{16}$$

If we apply the transformation (4) to the sum on the right of (16) it becomes

$$\sum_j\frac{km-k^2+3jk+3j-2jm+1}{(2j+1)(j+1)}\binom{j+1}{k-2j}\binom{m-k+3j}{2j}. \tag{17}$$

The creative telescoping algorithm with respect to k shows that the same recurrence

$$(k-2m)S(k)+(1-m+k)S(k+1)+(k+2)S(k+2)=0$$

is satisfied both by (17) and by the left side of (16). Since they agree for $k=0,1$, the identity is proved.

Problem 10223. (*Proposed in* 1992, *p.* 462; *solution in* 1997, *p.* 70)

For $p \in \mathbb{R}$, $q=1-p$, *and positive integers* n, *prove*

$$\sum_{k=n}^{2n-1}\binom{k-1}{n-1}\left(p^n q^{k-n}+p^{k-n}q^n\right)=1.$$

Write the sum as $S_n(p)+S_n(1-p)$ where $S_n(p)=\sum_{k=n}^{2n-1}\binom{k-1}{n-1}p^n(1-p)^{k-n}$. By algorithm ct

$$S_{n+1}(p)-S_n(p)=\frac{2p-1}{2}(p(1-p))^n\binom{2n}{n},$$

therefore

$$S_{n+1}(1-p)-S_n(1-p)=\frac{1-2p}{2}(p(1-p))^n\binom{2n}{n},$$

which implies that $S_n(p)+S_n(1-p)$ is constant. Evaluating the sum for $n=1$ completes the proof.

Problem 10229. (*Proposed in* 1992, *p.* 570; *solution in* 1994, *p.* 797)

Given that m *and* p *are integers with* $m \geq p \geq 1$, *evaluate*

$$\sum_{j=1}^{p}\binom{1/2}{m-j+1}\binom{1/2}{m+j}. \tag{18}$$

Let t_j denote the summand in (18). Gosper's algorithm finds that $t_j=s_{j+1}-s_j$ where $s_j=(j-1)(j+m)(2m-2j+1)t_j/(m(2m+1))$. Summing this recurrence on j from 1 to p gives the sum as $s_{p+1}-s_1$ which is

$$\binom{1/2}{m-p+1}\binom{1/2}{m+p}\frac{p(m-p+1)(2m+2p-1)}{m(2m+1)}.$$

Problem 10332. *Proposed in* 1993, *p.* 796; *solution in* 1996, *p.* 702)

If n *and* k *are integers with* $0 \leq k \leq n$, *prove that*

$$\binom{2n}{n+k}=\sum_j 2^{n-k-2j}\binom{n}{j}\binom{n-j}{j+k}.$$

If we divide the summand on the right by the claimed left side, $\binom{2n}{n+k}$, the WZ method proves the identity with a proof certificate of $4j(j+k)/((2j+k-n-1)(2n+1))$.

Problem 10357. (*Proposed in* 1994, *p.* 75; *solution in* 1997, *p.* 177)

Define integers $a_{m,n}$ *by*

$$\frac{1}{1-u-v+2uv}=\sum_{m,n=0}^{\infty}a_{m,n}u^m v^n.$$

Show that $(-1)^j a_{2j,2j+2}$ *is the Catalan number* $\binom{2j}{j}/(j+1)$.

We expand the function on the left into a power series in u and v, using first the geometric series, then the Binomial Theorem, and finally the derivatives of the geometric series:

$$\frac{1}{1-u-v+2uv} = \frac{1}{(1-v)\left(1-u\dfrac{1-2v}{1-v}\right)} = \frac{1}{1-v}\sum_{m=0}^{\infty}\left(\frac{1-2v}{1-v}\right)^m u^m$$

$$= \frac{1}{1-v}\sum_{m=0}^{\infty}\left(1-\frac{v}{1-v}\right)^m u^m = \sum_{m,k=0}^{\infty}\binom{m}{k}\frac{(-v)^k}{(1-v)^{k+1}}u^m$$

$$= \sum_{m,n,k=0}^{\infty}(-1)^k\binom{m}{k}\binom{n}{k}u^m v^n.$$

From this we see that $a_{m,n} = \Sigma_k(-1)^k\binom{m}{k}\binom{n}{k}$. Using Zeilberger's algorithm `ct` on $S(j) := (-1)^j a_{2j,2j+2} = \Sigma_k(-1)^{k+j}\binom{2j}{k}\binom{2j+2}{k}$ we obtain the recurrence

$$16(1+j)(1+2j)(9+4j)S(j) - 2(7+4j)(21+28j+8j^2)S(j+1)$$
$$+(3+j)(5+2j)(5+4j)S(j+2) = 0,$$

which is satisfied by $c_j = \binom{2j}{j}/(j+1)$. As $S(j) = c_j$ for $j = 0, 1$, the proof is complete.

Problem 10363. (*Proposed in* 1994, *p.* 175; *solution in* 1997, *p.* 179)

If m, n are integers satisfying $1 \le m \le n-1$, prove that

$$\binom{2n-m-1}{2n-2m-1} - \binom{n-1}{m} = \sum_k\sum_j\binom{k+j}{k}\binom{2n-m-2k-j-3}{2(n-m-k-1)}.$$

Let

$$S(k) = \sum_{j=0}^{m-1}\binom{k+j}{k}\binom{2n-m-2k-j-3}{2(n-m-k-1)}.$$

Algorithm `ct` finds the first-order recurrence

$$(2m-2n+k+1)S(k) + (2n-m-k-2)S(k+1) = 0,$$

for $0 \le k \le m-n-2$. From this, $S(k) = C(n,m)\binom{2m-2n+k}{k}/\binom{m-2n+k+1}{k}$, where

$$C(n,m) = S(0) = \sum_{j=0}^{m-1}\binom{2n-m-j-3}{2(n-m-1)} = \binom{2n-m-2}{m-1},$$

by Gosper's algorithm. Finally,

$$\sum_{k=0}^{n-m-1}S(k) = \binom{2n-m-2}{m-1}\sum_{k=0}^{n-m-1}\frac{\binom{2m-2n+k}{k}}{\binom{m-2n+k+1}{k}} = \binom{2n-m-1}{m} - \binom{n-1}{m},$$

by Gosper's algorithm again.

Problem 10375. (*Proposed in* 1994, *p.* 362; *solution in* 1997, *p.* 275)

Find the complete solution of the recurrence

$$U_{n+2} = 2(2n+3)^2U_{n+1} - 4(n+1)^2(2n+1)(2n+3)U_n, \quad \text{for } n \ge 0.$$

Petkovšek's algorithm `Hyper` finds that $(2n)!$ satisfies the recurrence. Then by reducing the order we find that $(2n)!H_n$ also satisfies it, where $H_n = 1 + 1/2 + \cdots + 1/n$ is the n-th harmonic number. Hence the complete solution is $U_n = (2n)!(C_1 + C_2H_n)$ where C_1 and C_2 are arbitrary constants.

Problem 10388. (*Proposed in* 1994, *p.* 474; *solution in* 1997, *p.* 459)

Find

$$\sum_{k=0}^{n} \binom{n}{k} \binom{\frac{n-3}{4} - \frac{k}{2} + p}{2p}$$

where n and p are positive integers.

The summand is not hypergeometric due to the non-integral coefficient of k. Denote the sum by $S(n, p)$, and write $S(n, p) = S_1(n, p) + S_2(n, p)$ where

$$S_1(n,p) = \sum_k \binom{n}{2k} \binom{\frac{n-3}{4} - k + p}{2p},$$

$$S_2(n,p) = \sum_k \binom{n}{2k+1} \binom{\frac{n-3}{4} - k - \frac{1}{2} + p}{2p}.$$

Zeilberger's algorithm `ct` finds that both sums satisfy the same recurrence with respect to p, viz.,

$$\begin{aligned}&(n-4p-7)(n-4p-5)(n-4p-3)(n-4p-1)S_i(n,p)\\ &-16(n-4p-7)(n-4p-5)(5n+4pn-8p^2-20p-14)S_i(n,p+1)\\ &+512(n-2p-3)(n-2p-4)(2p+3)(p+2)S_i(n,p+2) = 0, \quad \text{for } i \in \{1,2\}.\end{aligned} \tag{19}$$

Then $S(n, p)$ also satisfies (19). Algorithm `Hyper` finds the complete solution of this recurrence as

$$S(n,p) = C_1(n) \frac{\binom{\frac{n-3}{4}}{p}\binom{\frac{n-1}{4}}{p}}{(-4)^p \binom{-\frac{1}{2}}{p}\binom{\frac{n-2}{2}}{p}} + C_2(n) \frac{\binom{\frac{n-3}{4}}{p}\binom{\frac{n-1}{4}}{p}}{4^p \binom{\frac{n-1}{2}}{p}}.$$

The initial conditions $S(n, 0) = 2^n$ and

$$S(n,1) = \sum_{k=0}^{n} \binom{n}{k} \binom{\frac{n-3}{4} - \frac{k}{2} + 1}{2} = \frac{n-3}{32} 2^n,$$

can be found using algorithm `ct` again. From these $C_1(n) = 0$ and $C_2(n) = 2^n$, whence

$$S(n,p) = \frac{2^n}{4^p} \frac{\binom{\frac{n-3}{4}}{p}\binom{\frac{n-1}{4}}{p}}{\binom{\frac{n-1}{2}}{p}}.$$

Problem 10396. (*Proposed in* 1994, *p.* 681; *solution in* 1997, *p.* 570)

Let $\alpha > 0$ *and let* $\langle b_n : n \geq 1 \rangle$ *be defined recursively by* $b_1 = \alpha$, $b_2 = 3\alpha$,

$$b_{n+1} = (2n+1)b_n - (n^2 + \alpha^2)b_{n-1} \quad (n \geq 2). \tag{20}$$

Prove that $\langle b_n \rangle$ *contains infinitely many positive and infinitely many negative terms.*

Algorithm `Hyper` finds the complete solution of (20) as $C_1(1+\alpha i)(2+\alpha i)\cdots(n+\alpha i) + C_2(1-\alpha i)(2-\alpha i)\cdots(n-\alpha i)$ where $i^2=-1$ and C_1, C_2 are arbitrary constants. From the initial conditions, $C_1 = 1/(2i)$ and $C_2 = -1/(2i)$, so that

$$b_n = \Im(1+\alpha i)(2+\alpha i)\cdots(n+\alpha i).$$

Write $b_n = \Im z_n$, where $z_0 = 1$ and $z_n = (n+\alpha i)z_{n-1}$ (for $n \geq 1$). As $\lim_{n\to\infty}\arctan \alpha/n = 0$ but $\sum_{n=1}^{\infty}\arctan \alpha/n = \infty$, it is clear that the imaginary part of z_n is positive and negative infinitely often.

Problem 10403. (*Proposed in* 1994, *p.* 792; *solution in* 1997, *p.* 368)

Define a sequence $\langle y_n \rangle$ *recursively by* $y_0 = 1$, $y_1 = 3$ *and*

$$y_{n+1} = (2n+3)y_n - 2ny_{n-1} + 8n \tag{21}$$

for $n \geq 1$. *Find an asymptotic formula for* y_n.

Algorithm `Hyper` finds that $2^n n!$ satisfies the homogeneous part of (21). By reduction of order we obtain

$$y_n = 2^n n! \sum_{k=0}^{n} \frac{1+8\sum_{m=1}^{k-1} m}{2^k k!} = 2^n n! \sum_{k=0}^{n} \frac{1+4k(k-1)}{2^k k!}$$
$$= 2^{n+1} n! \sum_{k=0}^{n} \frac{1}{2^k k!} - 2n - 1,$$

which is asymptotic to $2^{n+1} n! \sqrt{e}$.

Problem 10424. (*Proposed in* 1995, *p.* 70; *solution in* 1997, *p.* 466)

Evaluate the sum

$$\sum_{0 \leq k \leq \frac{n}{3}} 2^k \frac{n}{n-k}\binom{n-k}{2k}. \tag{22}$$

Denote the sum in (22) by $S(n)$. The creative telescoping algorithm yields the constant-coefficient recurrence

$$S(n+3) - 2S(n+2) + S(n+1) - 2S(n) = 0 \quad (n \geq 1). \tag{23}$$

The roots of the characteristic polynomial are $2, \pm i$, and the solution of (23) satisfying $S(1) = S(2) = 1$ is

$$S(n) = 2^{n-1} - \sin\frac{(n-1)\pi}{2} = 2^{n-1} + \begin{cases} 0, & n \equiv 1 \pmod 2 \\ -1, & n \equiv 2 \pmod 4 \\ 1, & n \equiv 0 \pmod 4 \end{cases}.$$

Problem 10466. (*Proposed in* 1995, *p.* 654; *solution in* 1997, *p.* 575)

For $x \in \mathbb{C}$ *and* $n \in \mathbb{N}$, *prove the following identities between polynomials*:

$$(a) \qquad (-4)^n \sum_{j=0}^{n} \binom{x+\frac{1}{2}}{j}\binom{n-1-x}{2n-j} = \binom{2n}{n} \sum_{j=0}^{n} \binom{x+j}{2j}\binom{x-j}{2n-2j}$$

(*b*) *For all* $m \in \mathbb{N}$, *with* $0 \leq m \leq 2n$, *generalize* (*a*) *to*

$$(-4)^n \sum_{j=0}^{n} \binom{x+\frac{1}{2}}{j}\binom{n-1-x}{2n-j} = \binom{2n}{n} \sum_{j=-\lfloor \frac{m}{2} \rfloor}^{n-\lfloor \frac{m}{2} \rfloor} \binom{x+j}{2j+m}\binom{x-j}{2n-m-2j}$$

(a) Denote the sums on the left and right by $S(n)$ and $T(n)$, respectively. Algorithm ct finds recurrence relations

$$(1+2n)S(n) - 2(1+n)S(n+1) = \frac{(1+3n-2x)(1+2x)}{1+n}\binom{2x}{2n}\binom{2n}{n} \tag{24}$$

and

$$\begin{aligned} &2(1+n)(2+n)^2(1+3n-2x)T(n+2) \\ &-(1+n)(6+41n+67n^2+30n^3 \\ &\qquad -x(28+98n+68n^2-40x-56nx+16x^2))T(n+1) \\ &+2(1+2n)(4+3n-2x)(x-n)(2x-2n-1)T(n) = 0. \end{aligned}$$

The latter recurrence turns out to be the homogenization of the former, so $S(n)$ and $T(n)$ satisfy the same recurrence of order 2. As they agree for $n = 0, 1$, they are identical. As a bonus, from (24) we can express $S(n)$ in terms of an "indefinite" sum in which the summand does not depend on n:

$$S(n) = \frac{\binom{2n}{n}}{4^n}\left(1 - (2x+1)\sum_{k=0}^{n-1} 4^k \frac{3k+1-2x}{(k+1)(2k+1)}\binom{2x}{2k}\right).$$

(b) Denote the sum on the right by $U(n, m)$. Algorithm ct with respect to m finds the recurrence

$$(2n-m-1)U(n,m+2) - 2nU(n,m+1) + (m+1)U(n,m) = 0. \tag{25}$$

For $U(n,1) = \binom{2n}{n}\sum_{j=0}^{n-1}\binom{x+j}{2j+1}\binom{x-j}{2n-1-2j}$, algorithm ct finds the same recurrence as for $T(n)$ (for $n \ge 1$). As they agree for $n = 1, 2$, they agree for all $n \ge 1$. So $U(n,1) = T(n)$ $= U(n,0)$ (for $n \ge 1$). It follows from (25) that $U(n,m) = T(n) = S(n)$.

Problem 10473. (*Proposed in* 1995, *p.* 745; *solution in* 1997, *p.* 371)

Prove that there are infinitely many positive integers m such that

$$\frac{1}{5\cdot 2^m}\sum_{k=0}^{m}\binom{2m+1}{2k}3^k \tag{26}$$

is an odd integer.

Denote (26) by $S(m)$. Algorithm ct yields the constant-coefficient recurrence

$$S(m+2) - 4S(m+1) + S(m) = 0. \tag{27}$$

The sequence $T(m) = 5S(m)$ satisfies (27) as well and starts out as $\langle 1, 5, \dots \rangle$, hence it is integral. Let $T_5(m) = T(m) \bmod 5$ and $T_2(m) = T(m) \bmod 2$. Using (27) mod 5 and mod 2, respectively, we see that $T_5 = \langle 1,0,4,1,0,4,\dots\rangle$ and $T_2 = \langle 1,1,1,\dots\rangle$, so that $S(3k+1)$ $= T(3k+1)/5$ is an odd integer for all $k \ge 0$.

Problem 10494. (*Proposed in* 1996, *p.* 74; *solution in* 1997, *p.* 371)

For each positive integer n, evaluate the sum

$$\sum_{k=0}^{2n}(-1)^k\binom{4n}{2k}\bigg/\binom{2n}{k}. \tag{28}$$

Let t_k denote the summand in (28). Gosper's algorithm finds that $t_k = s_{k+1} - s_k$ where $s_k = (2k-1)t_k/(2(1-2n))$. Summing on k from 0 to $2n-1$ gives the sum as $s_{2n} - s_0 +$ $t_{2n} = 1/(1-2n)$.

6. CONCLUSION.

Quite often, MONTHLY problems require evaluation of a single or double sum in closed form, or a proof of equality of two such sums. When the summand involves binomial coefficients, factorials, products of rational functions,

and exponential functions with constant base, there are very good chances that such a problem can be solved automatically by Gosper's algorithm or by its generalizations: the WZ method, Zeilberger's algorithm `ct`, and algorithm `Hyper`. Although Gosper's algorithm is now over 19 years old (see the quotation on the title page!), it seems that it is not as widely known as it deserves to be.

To help spread the word, we surveyed MONTHLY problems that have appeared since the publication of Gosper's algorithm in 1978. We have presented here a selection of those on which these methods are successful. For a similar list of earlier problems, see the Web site
`http://www.math.temple.edu/~zeilberg/Monthly.html.`

REFERENCES

1. Sister Mary Celine Fasenmyer, *Some Generalized Hypergeometric Polynomials*, Ph.D. dissertation, Univ. of Michigan, 1945.
2. R. W. Gosper, Jr., Indefinite hypergeometric sums in MACSYMA, *Proc. MACSYMA Users Conference*, Berkeley CA, 1977, 237–252.
3. R. W. Gosper, Jr., Decision procedure for indefinite hypergeometric summation, *Proc. Natl. Acad. Sci. USA* **75** (1978) 40–42.
4. P. Paule, *Solution of a Séminaire homework example* (*28th SLC*), RISC-Linz Report Series No. 92–59, Linz 1992.
5. P. Paule and M. Schorn, A *Mathematica* version of Zeilberger's algorithm for proving binomial coefficient identities, *J. Symb. Comput.* **20** (1995) 673–698.
6. M. Petkovšek, Hypergeometric solutions of linear recurrences with polynomial coefficients, *J. Symb. Comput.* **14** (1992) 243–264.
7. M. Petkovšek and H. S. Wilf, When can the sum of $(1/p)$th of the binomial coefficients have closed form?, *Electronic J. Combinatorics*, to appear.
8. M. Petkovšek, H. S. Wilf and D. Zeilberger, $A = B$, A K Peters, Ltd., Wellesley, MA, 1996.
9. H. S. Wilf and D. Zeilberger, Rational functions certify combinatorial identities, *J. Amer. Math. Soc.* **3** (1990) 147–158.
10. H. S. Wilf and D. Zeilberger, An algorithmic proof theory for hypergeometric (ordinary and "q") multisum/integral identities, *Inv. Math.* **108** (1992) 575–633.
11. D. Zeilberger, The method of creative telescoping, *J. Symb. Comput.* **11** (1991) 195–204.

RISC Linz
Johannes Kepler University
Linz, Austria
Istvan.Nemes@risc.uni-linz.ac.at

University of Ljubljana
Ljubljana, Slovenia
Marko.Petkovsek@mat.uni-lj.si

University of Pennsylvania
Philadelphia, PA, USA
wilf@math.upenn.edu

Temple University
Philadelphia, PA, USA
zeilberg@euclid.math.temple.edu

Simultaneously Symmetric Functions

Lawrence W. Baggett, Herbert A. Medina, and Kathy D. Merrill

1. INTRODUCTION. Symmetry properties of functions are traditional tools used to simplify the processes of both graphing and integration. In addition, a knowledge of the symmetry properties of initial conditions and boundary data provides qualitative information about the time evolution of solutions to differential equations [3]. The most immediately useful types of symmetry are contained in the notions of even and odd functions. We make these notions precise with the following definition.

Definition. A real-valued function w is said to be *even* on an interval $[a, b]$ if $w(a + x) = w(b - x)$ for all $x \in [0, b - a]$. The function is *odd* on $[a, b]$ if $w(a + x) = -w(b - x)$.

Sometimes we will call a function *even* or *odd* if the corresponding equality holds almost everywhere. Since such precision is not critical to our results, we will frequently omit the phrase almost everywhere in this and other contexts. In particular, this gives us license to ignore the endpoints of intervals.

Any function w on $[a, b]$ can be broken down uniquely into the sum of an even function and an odd function by writing

$$w(x) = \frac{w(x) + w(a + b - x)}{2} + \frac{w(x) - w(a + b - x)}{2}, \quad x \in [a, b].$$

This representation is unique since only the identically 0 function is both even and odd on $[a, b]$. Thus we can speak without ambiguity of the *even part* w_e and the *odd part* w_o of a function w on $[a, b]$.

Our ability to break any function on $[a, b]$ into its even and odd parts leads to breaking real function spaces on $[a, b]$ down into the even functions on $[a, b]$ and the odd functions on $[a, b]$. Since any even function is orthogonal to any odd function under the inner product $(v, w) = \int_a^b v(x)w(x)\,dx$, this breaks the real inner product space $L^2[a, b]$ into a direct sum of orthogonal complements. This process is implicit in many of the traditional bases used for $L^2[a, b]$. For example, the Fourier basis for $L^2[-\pi, \pi]$, consists of the (even) cosine functions and the (odd) sine functions. The simplest polynomial basis for functions on $[-a, a]$, $\{1, x, x^2, \dots\}$ is traditionally broken down into the odd powers, which are odd functions on $[-a, a]$, and the even powers, which are even functions. As a final example, the Legendre polynomials, which arise from applying the Gram-Schmidt process to the powers of x restricted to any interval $[a, b]$, also have the property that the even degree members are even functions and the odd degree members are odd functions on that interval.

In this paper, we will be concerned with the possibility of simultaneous symmetry on several intervals. That is, we cut an interval $[a, b]$ into two pieces $[a, c]$ and $[c, b]$ and ask two questions: First, under what conditions can a function be simultaneously even (or simultaneously odd) on the whole interval and on the two pieces? Second, under what conditions can a function be simultaneously even (or

simultaneously odd) on two of these three intervals? Since we can think of $L^2[a, c]$ and $L^2[c, b]$ as being contained in $L^2[a, b]$, we are in effect asking about the intersection of the three even or three odd subspaces.

One way to picture these questions physically is by imagining standing waves on a string whose ends are fixed at the points a and b. If the midpoint of the string is constrained to be a node, the resulting wave pattern is forced to be odd on the whole string. Now choose two other points $a < p < q < b$ so that the distance between them is the sum of the distances to their respective ends: $q - p = p - a + b - q$. These two points are then the midpoints of two subintervals, which form a subdivision of our interval. If we constrain the string to force these points to be nodes as well, we force the wave pattern to be odd on the two subintervals of which they are the centers. Is this possible no matter how we choose our points p and q? In this paper we ask first the analogous question about functions that are allowed to be more complicated than standing waves. We then relax our requirement to demand only that two of the three midpoints be nodes.

These questions might be mere curiosities except that their answers have strong connections to the field of ergodic theory, which studies the long-term behavior of dynamical systems. A mapping T from a measure space X into itself is called *ergodic* if it mixes the space X up effectively in the following sense: any measurable function $w\colon X \to \mathbb{R}$ that satisfies $w(Tx) = w(x)$ for almost every $x \in X$ (we say that such a w is *invariant* under T), must be constant. The answer to our first question is related to the ergodicity of a natural class of mappings from the interval $[a, b]$ into itself that come from bending $[a, b]$ into a circle by identifying a and b, and then defining T to be rotation through a fixed angle. These rotations are ergodic exactly when the rotation is through an angle irrationally related to 2π; on $[a, b]$, this happens when the rotation is induced by a number irrationally related to $b - a$ ([**4**], [**5**]). In Section 2, we see that nonconstant functions exist that are simultaneously even (or odd) on all three intervals if and only if the rotation that takes a to the division point c is not ergodic. Further, when nonconstant simultaneously symmetric functions do exist, they are precisely the invariant functions that cause the rotation to be non-ergodic. Thus, simultaneous symmetry on three intervals gives us a simple geometric way of picturing ergodicity of rotations.

In Section 3, we see that simultaneous symmetry on two of the three intervals has a similar interpretation, which depends on a related functional equation from ergodic theory. A real-valued measurable function v is called a *coboundary* for a mapping T if there exists a real-valued measurable function w such that $v(x) = w(x) - w(Tx)$ for almost every x. The difficult problem of determining which functions are coboundaries is important in the theory of invariant measures and of generalized eigenvalues. It is also necessary in understanding a key class of examples in ergodic theory called skew products ([**4**], [**5**]). By writing the coboundary equation in the form $w(Tx) = w(x) - v(x)$, we can think of the w's that solve it as being "nearly invariant" under T, with the corresponding v's being the obstructions to actual invariance. We show that when simultaneous symmetry on all three intervals is impossible, the obstructions to it are precisely the same as these obstructions to invariance, that is, coboundaries. We also find that while simultaneous symmetry on two of the three intervals is always possible, these doubly symmetric functions are constrained to be the corresponding "nearly invariant" functions w. Thus, simultaneous symmetry on two of the three intervals gives a geometric interpretation of solutions to the coboundary equation for irrational rotations.

2. SIMULTANEOUS SYMMETRY ON THREE INTERVALS. Because differences of scaling and translation are easily dealt with, we focus on the case $[a, b] = [0, 1]$. Thus we ask the following question:

Question. *Under what conditions on the real number θ, $0 < \theta < 1$, and on the function w on $[0, 1]$, can w be simultaneously even (or odd) on the three intervals $[0, \theta]$, $[\theta, 1]$, and $[0, 1]$?*

If θ is a rational number p/q, then we can easily construct functions with this property of simultaneous symmetry. We simply take any symmetric function on $[0, 1/q]$ and repeatedly translate it by $1/q$ to cover all of $[0, 1]$. For example, the function $f(x) = \cos 10\pi x$ is even on the intervals $[0, 1]$, $[0, 2/5]$, and $[2/5, 1]$. (See Figure 1.)

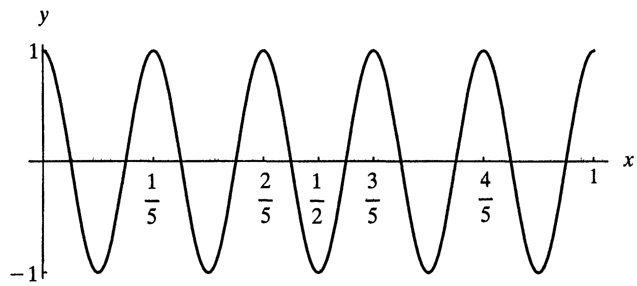


Figure 1. Graph of $f(x) = \cos 10\pi x$.

In fact, such examples are the only functions that are simultaneously symmetric on $[0, 1]$, $[0, \theta]$, and $[\theta, 1]$, and they are possible only if θ is rational. This is shown in the following theorem.

Theorem 1. *Let $\theta \in (0, 1)$ be given and suppose that a function w is simultaneously even (or simultaneously odd) on the three intervals $[0, 1]$, $[0, \theta]$, and $[\theta, 1]$. Then w is invariant under rotation by θ. In particular, if θ is rational and $\theta = p/q$ (in lowest terms), then w is the periodic extension to $[0, 1]$ of an even (odd) function on $[0, 1/q]$; if θ is irrational, then w is a constant.*

Proof: Extend w periodically from $[0, 1)$ to $\mathbb{R}$, so that w can be regarded as a function on the circle $\mathbb{R}/\mathbb{Z}$, with rotation represented by addition mod 1.

If w is even on all three intervals, then w satisfies the three conditions:

$$w(x) = w(1 - x) \quad \text{for } 0 \le x \le 1, \tag{1}$$

$$w(x) = w(\theta - x) \quad \text{for } 0 \le x \le \theta, \tag{2}$$

$$w(\theta + x) = w(1 - x) \quad \text{for } 0 \le x \le 1 - \theta. \tag{3}$$

Combining (1) and (2) gives $w(x) = w(x + \theta - 1) = w(x + \theta)$ for $1 - \theta \le x \le 1$. Combining (1) and (3) gives the same equality for $0 \le x \le 1 - \theta$. This establishes that w is invariant under translation by θ on $[0, 1]$. Thus, if $\theta = p/q$, we see by iterating this equation that w is determined by translating its restriction to $[0, 1/q]$.

Combining this with (1) then shows that w is even on $[0, 1/q]$. On the other hand, if θ is irrational, ergodicity of translation by θ on $[0,1]$ (rotation on $\mathbb{R}/\mathbb{Z}$) ensures that w must be constant. The same proof, with the insertion of the appropriate minus signs, establishes the result when w is simultaneously odd on the three intervals. In the latter case, if θ is irrational, w must be 0. ∎

Theorem 1 establishes the close connection between simultaneous symmetry and ergodicity of rotations claimed in the Introduction. Indeed, the proof works by showing that the simultaneously symmetric functions are precisely the functions that are symmetric on the whole interval and invariant under the rotation that takes 0 to θ.

For our standing wave example, Theorem 1 shows that if θ is irrational, we cannot create standing waves with the three midpoints as nodes. On the other hand, if $\theta = p/q$, we see that linear combinations of $\sin 2\pi nx$, for n's that are multiples of q, are odd on all three intervals, and linear combinations of $\cos 2\pi nx$ are even on all three. We can get all L^2 functions that are simultaneously symmetric on all three intervals by building Fourier series out of these symmetric pieces.

In the next section, we relax our requirement to symmetry on two of the three intervals.

3. SIMULTANEOUS SYMMETRY ON TWO INTERVALS. In the preceding section, we saw that simultaneous symmetry on $[0,1]$, $[0,\theta]$, and $[\theta,1]$ implies a kind of invariance under rotation that severely restricts the type of functions that are possible. We will see in this section that simultaneous symmetry on two of these three intervals still gives a "near invariance," and corresponding restrictions. Our first use of this is in the following theorem.

Theorem 2. *Let θ be any number, rational or irrational, in $(0,1)$, and suppose p is a polynomial on $[0,1]$ that is even (odd) on two of the three intervals $[0,1]$, $[0,\theta]$, and $[\theta,1]$. Then p is a constant polynomial.*

Proof: We treat the simultaneously even case; the argument for the odd case is analogous. The assumption implies that at least two of the three equations (1), (2), and (3) in the proof of Theorem 1 must hold for p. If (1) and (3) hold, it follows as in that proof that $p(x) = p(x+\theta)$ for an infinite number of x's, and hence for all complex numbers x. If (1) and (2) hold, then (since here we do not extend p periodically, and hence cannot assume that addition in the argument can be taken mod 1) $p(x) = p(x+\theta-1)$ for an infinite number of x's. In either case, p has a nonzero period and hence is constant.

If (2) and (3) hold, then (2) ensures that $p(x) = p(\theta - x)$ for an infinite number of x's, and hence for all complex numbers x. Therefore, $p(-x) = p(\theta + x)$, and combining this with (3) ensures that $p(-x) = p(1-x)$. This again leads to the conclusion that p is constant. ∎

The preceding proof shows that any real-analytic function that is simultaneously symmetric on two of the three intervals extends to an entire function with a nontrivial period. This periodicity on $\mathbb{R}$ is similar to the invariance on $\mathbb{R}/\mathbb{Z}$ required of a function that is simultaneously symmetric on all three intervals. Since trigonometric functions with irrational periods are periodic on $\mathbb{R}$ but not invariant on $\mathbb{R}/\mathbb{Z}$, this suggests that they might provide a source of examples of functions

that are simultaneously symmetric on two of the three intervals. We will see later that this is indeed the case, but first we develop some helpful machinery.

There are three possible pairs of intervals to consider, and to distinguish them more clearly, we adopt a convention that $\theta \le 1/2$, so that $[0, \theta]$ is the smaller of the two subintervals.

The easiest case to picture is the one in which the function w is symmetric on the two subintervals $[0, \theta]$ and $[\theta, 1]$. In this case, it is easy to see that any symmetric function on $[0, \theta]$ can be pasted together with any symmetric function on $[\theta, 1]$ to give a function on $[0, 1]$ of the desired type. Figure 2 illustrates an example of a function that is even on $[0, \theta]$ and $[\theta, 1]$ with $\theta = \sqrt{2}/5 \approx 0.282843$.

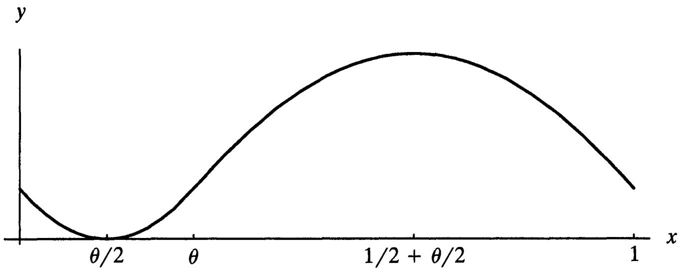


Figure 2. Graph of a function that is even on $[0, \theta]$ and $[\theta, 1]$.

It requires some work to discover what characteristics a function that is symmetric on the two subintervals must have on $[0, 1]$. We will return to this question in a moment. First, we show that the other two cases of simultaneous symmetry on two of the intervals are just scaled down versions of this one.

Consider the case of a function that is simultaneously even (odd) on $[0, 1]$ and $[0, \theta]$. Since the function is even (odd) on $[0, 1]$, it must also be so on any subinterval centered at $\frac{1}{2}$, in particular on $[\theta, 1 - \theta]$. Thus, if we restrict our attention to the interval $[0, 1 - \theta]$, we see again the first type of simultaneous symmetry on two intervals. That is, we think of $[0, \theta]$ and $[\theta, 1 - \theta]$ as a subdivision of $[0, 1 - \theta]$, and see that what we are seeking is a function that is even (odd) on the two pieces of the subdivision. Further, this restriction to $[0, 1 - \theta]$ completely determines the function on $[0, 1]$ since symmetry on $[0, 1]$ forces the function on $[1 - \theta, 1]$ to be the mirror image (or minus the mirror image) of the function on $[0, \theta]$. Figure 3 is the graph of a function that is odd on $[0, \theta]$ and $[0, 1]$ with $\theta = \sqrt{5}/10 \approx 0.223607$.

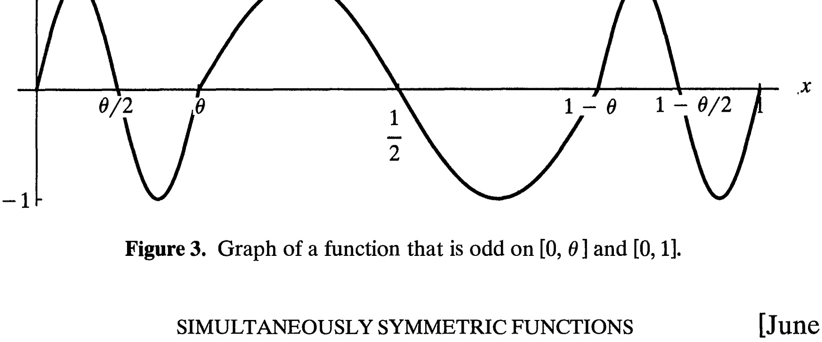


Figure 3. Graph of a function that is odd on $[0, \theta]$ and $[0, 1]$.

The final case of simultaneous symmetry on two of the three intervals is that of functions that are even (odd) on the whole interval $[0,1]$ and on the larger of the two pieces, $[\theta,1]$. To see that this case again reduces to a scaled down version of the first case, construct the following picture (for simplicity we consider the simultaneously even case): Draw an arrow from 0 to θ to represent the function's evolution as x goes from 0 to θ. Then, using the symmetry on $[0,1]$, draw an arrow from 1 back to $1-\theta$ to indicate the behavior of the function on $[1-\theta,1]$. Now use symmetry on $[\theta,1]$ to draw that same arrow from θ to 2θ. Then use symmetry on $[0,1]$ to draw it from $1-\theta$ to $1-2\theta$. Repeat this process until no more multiples of θ will fit. The function in Figure 4 is an example of such a construction with $\theta=\pi/15\approx 0.20944$.

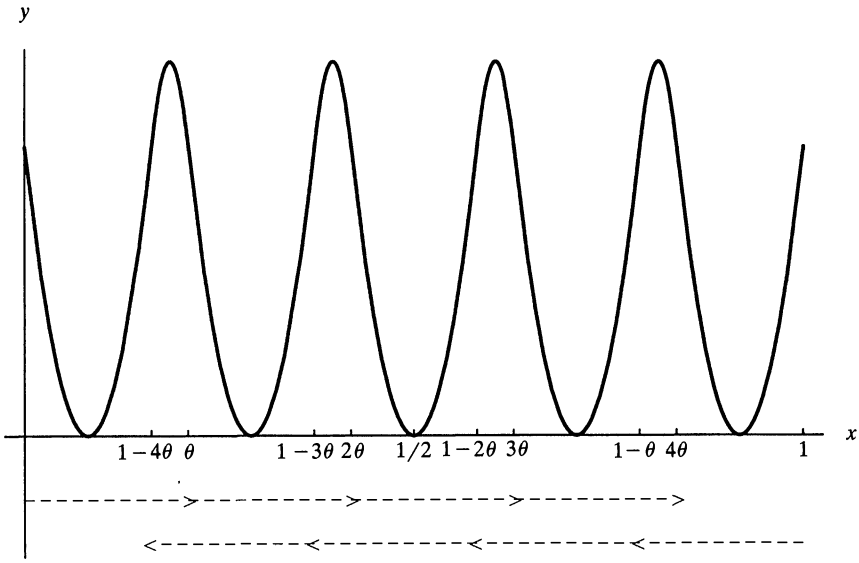


Figure 4. Graph of a function that is even on $[\theta,1]$ and $[0,1]$.

The resulting picture shows the nearly periodic behavior required of any such function. Also, the function must be even on any subinterval where two tails or two heads of arrows overlap in opposite directions. These results are contained in the following theorem, which we state and prove for even functions; the odd analogue is clear.

Theorem 3. *Let θ be an irrational number in $(0,1)$, and let q denote the greatest integer in $1/\theta$. A function w is simultaneously even on $[0,1]$ and $[\theta,1]$ if and only if w is the extension via the equation $w(x)=w(x+\theta)$ of a function on $[0,\theta]$ that is simultaneously even on $[0,1-q\theta]$ and $[1-q\theta,\theta]$.*

Proof: As before, we first extend w periodically from $[0,1)$ to $\mathbb{R}$. If w is simultaneously even on $[0,1]$ and $[\theta,1]$, then we have conditions (1) and (3) as in the proof of Theorem 1, again yielding that $w(x)=w(x+\theta)$ for $x\in[0,1-\theta]$. Iterating this equation gives $w(x)=w(x+q\theta)$ for $x\in[0,1-q\theta]$. Combining this with (1) gives the evenness of w on $[0,1-q\theta]$. To establish the evenness on $[1-q\theta,\theta]$, stop the iteration one step sooner, with $w(x)=w(x+(q-1)\theta)$ for $x\in[0,1-(q-1)\theta]$. Thus $w(\theta-x)=w(q\theta-x)=w(1-q\theta+x)$ for $x\in[0,(q+1)\theta-1]$.

Conversely, given an $x \in [0,1]$, let k denote the greatest integer in x/θ. Then $x - k\theta \in [0, \theta]$, and the hypothesis implies that $w(x) = w(x - k\theta)$. If $x - k\theta \in [0, 1 - q\theta]$, then by symmetry on that interval we have

$$w(x) = w(x - k\theta) = w(1 - q\theta - (x - k\theta)) = w(1 - x + (k - q)\theta) = w(1 - x).$$

Thus also $w(x + \theta) = w(1 - x)$. The proof for the case $x - k\theta \in [1 - q\theta, \theta]$ is similar. ∎

Since we have seen that all three cases of functions even on two of the three intervals $[0,1]$, $[0, \theta]$, and $[\theta, 1]$ reduce to a scaled-down version of a function even on the two subintervals, we focus now on that case. We consider the question of what types of functions are simultaneously even (or simultaneously odd) on $[0, \theta]$ and $[\theta, 1]$. One way to think about this question is to ask what the possible obstructions are to simultaneous symmetry on all three intervals. That is, we ask what the odd part of a function can be on $[0, 1]$ if that function is even on both $[0, \theta]$ and $[\theta, 1]$.

The answer turns out to be related to the coboundary equation described in the Introduction. In our context of rotation on $[0, 1]$, this equation takes the following form: we say that a measurable function v on $[0, 1]$ is a *coboundary* for the number θ if there exists a measurable function w such that $v(x) = w(x) - w(x + \theta)$ for almost every $x \in [0, 1]$; addition in the argument is taken mod 1. The function w is called a *transfer* or *cobounding* function for v. Note that for a fixed v, any two transfer functions must differ by an invariant function. Because of ergodicity for an irrational θ, this implies that the transfer function is unique up to an additive constant. However, for a rational θ, the transfer functions for a fixed v can come in many shapes.

We state and prove Theorem 4 for odd functions and note that it has an even analogue.

Theorem 4. *Fix $\theta \in [0, 1]$. If an odd function v is the odd part on $[0, 1]$ of a function w that is even on both $[0, \theta]$ and $[\theta, 1]$, then v is a coboundary for θ with transfer function $w/2$. Conversely, if v is a coboundary for θ, then v has a transfer function w that is even on both $[0, \theta]$ and $[\theta, 1]$ and whose odd part on $[0, 1]$ is $v/2$.*

Proof: Suppose first that v is the odd part on $[0, 1]$ of a function w that is even on both $[0, \theta]$ and $[\theta, 1]$. Extend v periodically from $[0, 1)$ to $\mathbb{R}$. Then (2) and (3) hold for w, as in the proof of Theorem 1, while (1) is replaced by

$$w(1 - x) = w(x) - 2v(x) \quad \text{for } 0 \leq x \leq 1. \tag{1'}$$

Combining first (1′) and (2) and then (1′) and (3) as in the proof of Theorem 1, we deduce that $w(x) - w(x + \theta) = 2v(x)$.

Conversely, suppose now that the odd function v is a coboundary for θ with transfer function w, so that $w(x) - w(x + \theta) = v(x)$. Then, since v is odd on $[0, 1]$, we have

$$w(x) - w(x + \theta) = -w(1 - x) + w(1 - x + \theta),$$

and thus

$$w(x) - w(1 - x + \theta) = w(x + \theta) - w(1 - x),$$

so that the function $u(x) = w(x) - w(1 - x + \theta)$ is invariant under translation by θ. Thus $\tilde{w} = w - u/2$ is also a transfer function for v. For $x \in [0, \theta]$, $u/2$ is the odd part of the restriction to $[0, \theta]$ of w, implying that the odd part of $\tilde{w}$ restricted

to $[0, \theta]$ is 0. This proves that $\tilde{w}$ is even on $[0, \theta]$. And, for $x \in [0, 1-\theta]$, the translate of $u/2$ by θ, which equals $u/2$, is the odd part of the restriction to $[\theta, 1]$ of w, showing that $\tilde{w}$ is also even there.

We finish the proof by noting that the odd part of $\tilde{w}$ is

$$\frac{\tilde{w}(x) - \tilde{w}(1-x)}{2} = \frac{\tilde{w}(x) - \tilde{w}(x+\theta)}{2} = \frac{v(x)}{2}. \qquad \blacksquare$$

By showing that the problem of simultaneous symmetry is equivalent to the problem of finding coboundaries for irrational θ, we have found a simple geometric description for that difficult problem. We have also acquired for the simultaneous symmetry problem all the machinery and results developed in the coboundary context. The following coboundary result is particularly useful [**1**].

Theorem 5. *A function $v \in L^2$ is a coboundary for an irrational θ with L^2 transfer function if and only if its sequence of Fourier coefficients, $\{\hat{v}(n)\}$, satisfies*

$$\sum_{n \neq 0} \left| \frac{\hat{v}(n)}{(1 - e^{2\pi i n\theta})} \right|^2 < \infty.$$

Proof: A measurable function $v \in L^2$ is a coboundary with an L^2 transfer function w if and only if the elements of the sequence $\{\hat{w}(n)\}$, which solve the equation $\Sigma \hat{v}(n) e^{2\pi i n x} = \Sigma \hat{w}(n) e^{2\pi i n x}(1 - e^{2\pi i n\theta})$, are the Fourier coefficients of an L^2 function; i.e., they are in l^2. Solving for $\hat{w}(n)$ yields the condition in the statement of the theorem. $\blacksquare$

In particular, we can apply Theorem 5 to see that trigonometric polynomials are all coboundaries. For example, $\sin 2\pi x$ is an odd function that is a coboundary, so by Theorem 4 it is the odd part of a function w that is even on $[0, \theta]$ and $[\theta, 1]$. Further, we know that the function w is twice the transfer function of $\sin 2\pi x$, so we can solve for w using Fourier coefficients as in Theorem 5. We find that

$$w(x) = \sin 2\pi x + \frac{\sin 2\pi\theta}{1 - \cos 2\pi\theta} \cos 2\pi x$$

is even on $[0, \theta]$ and $[\theta, 1]$. For example, if $\theta = \sqrt{3}/4 \approx 0.433013$, Figure 5 shows the graphs of $v(x) = 2 \sin 2\pi x = w(x) - w(x+\theta)$ and w.

A similar example can easily be built with $\cos 2\pi x$ to give a simple function that consists of standing waves constrained to have nodes at the midpoints of each of

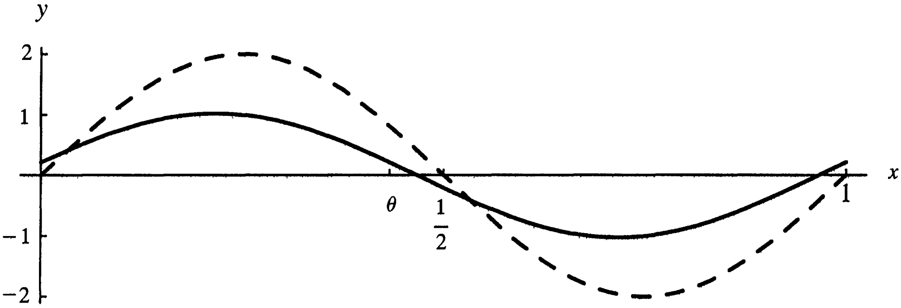


Figure 5. Graphs of $v(x) = 2\sin 2\pi x$(- -), $w(x) = \sin 2\pi x + \dfrac{\sin 2\pi\theta}{1 - \cos 2\pi\theta} \cos 2\pi x$.

the two subintervals resulting from an irrational subdivision. Theorem 5 shows that we can build more complicated functions with this property by using finite linear combinations of $\cos 2\pi kx$. Infinite combinations are more problematic, with a tradeoff between how smooth the resulting v is and how well θ can be approximated by rationals. (e.g., see [**2**].)

ACKNOWLEDGMENT. We wish to thank Professor Barbara Whitten for suggesting the physical interpretation of standing waves for our mathematical results.

REFERENCES

1. L. Baggett and K. D. Merrill, Representations of the Mautner group and cocycles of an irrational rotation, *Michigan Math. J.* 33 (1986), 221–229.
2. L. Baggett and K. D. Merrill, Smooth cocycles for an irrational rotation, *Israel J. of Math.*, 79 (1992), 281–288.
3. R. V. Churchill and J. W. Brown, *Fourier Series and Boundary Value Problems* (4th ed.), McGraw-Hill, New York, 1987.
4. P. R. Halmos, *Lectures on Ergodic Theory*, Chelsea Publishing Company, New York, 1956.
5. K. Petersen, *Ergodic Theory*, Cambridge University Press, Cambridge, 1983.

Department of Mathematics
University of Colorado
Boulder, Colorado 80309
baggett@euclid.colorado.edu

Department of Mathematics
Loyola Marymount University
Los Angeles, California 90045
hmedina@lmumail.lmu.edu

Department of Mathematics
The Colorado College
Colorado Springs, Colorado 80903
kmerrill@cc.colorado.edu

From the MONTHLY Fifty Years Ago . . .

Computing machines, by Myron Tribus, Lecturer in Engineering, University of California at Los Angeles, introduced by Professor Clifford Bell.

The recent developments in computing machines were described and some details of the unique features of the new machines were given. It was suggested that the machines of the future may be capable of performing 20,000 to 40,000 computations per second. (p. 511)

It is truly amazing what powers the vacuum tube gives to mankind . . . In the next few years it will probably supply jobs to millions of men in completely new industries, and will in thousands of ways increase our safety, comfort, and material wealth, and happiness. (p. 436)

The Harvard Sequence Controlled Calculator, built by the International Business Machine Corp. under the guidance of Prof. Howard A. Aiken, is a major engineering project. It is an assemblage of calculating and control elements, mounted on racks 8 feet high and totalling 63 feet in length, and weighing about 5 tons. A 4-horsepower motor furnishes its mechanical power through a network of shafts and gears. (p. 57)

MONTHLY 54 (1947)

Prime-Producing Quadratics

R. A. Mollin

Dedicated to the memory of Daniel Shanks

1. INTRODUCTION. From the recreational mathematician to the research mathematician, prime producing quadratic polynomials have held a longstanding fascination. These polynomials have been ubiquitous in the literature for centuries, but quite often they appear merely as curiosities, or with explanations that are incomplete. This article is intended to explain the reasons behind this prime production to anyone from the uninitiated reader to the expert. The reasons are given in terms of class group structures of quadratic fields, which the reader is brought to understand via a development from the basics. We take the reader from the fundamental idea of a quadratic field, through the arithmetic of ideals therein. Furthermore, complete lists are given of quadratic polynomials, having negative discriminant, that generate consecutive, distinct primes for an initial range, and these lists are shown to be complete under the assumption of a suitable Riemann hypothesis. Attendant topics are also discussed in detail, such as the "density" of primes produced by such polynomials and the current record holder in that regard, as well as material buried at various depths throughout the literature.

2. COMPLEX PRIME-PRODUCERS. The most celebrated of the quadratic prime-producing polynomials is $x^2 - x + 41$, discovered by Euler in 1772 [**9**]. This polynomial is prime for integers $x = 1, 2, 3, \ldots, 40$. Similarly, in 1798, Legendre [**15**] observed that the polynomial

$$f(x) = x^2 + x + 41$$

is prime for all integers $x = 0, 1, \ldots, 39$. It is the latter that has become known as *Euler's polynomial* (for example, see [**24**]). In any case, the prime-producing capacity of these polynomials has less to do with their specific form than it does with their discriminant,[1] which is -163. We can find numerous polynomials of discriminant -163 that generate consecutive prime values for at least 40 integer values of x simply by translating the Euler polynomial. For example, we can produce an infinite family of polynomials that generate forty consecutive, distinct prime values. Consider the following polynomial, for each $n \in \mathbb{Z}$, achieved from $f(x)$ via $x \to 3x - 39 - 3n$:

$$g_n(x) = 9x^2 - (18n + 231)x + 9n^2 + 231n + 1523,$$

[1]Recall that the *discriminant* of $f(x) = ax^2 + bx + c$ is $b^2 - 4ac$. This is Lagrange's notion of a discriminant, which differs from that of Gauss, who considered forms of the type $ax^2 + 2bxy + cy^2$, and defined the *determinant* as $b^2 - ac$, which has become known as *Gauss's discriminant.* See [**19**, Appendix E, pp. 347–354] for a detailed explanation of the relationship between forms and ideals, including clarification of problems and correction of some errors that have crept into the literature over the years.

which produces forty consecutive, distinct primes for the values:

$$g_n(x+n) = \begin{cases} f(38-3x) & \text{for } x = 0,1,\ldots,12, \\ f(3x-39) & \text{for } x = 13,14,\ldots,39. \end{cases}$$

These polynomials merely pick the prime values of $f(x)$, spaced three apart, thereby reflecting the discriminant of $g_n(x)$, namely $-3^2 \cdot 163$. For instance, if $n = 0$, then $g_0(x) = 9x^2 - 231x + 1523$ is prime for $x = 0, 1, \ldots, 39$ (the special case discovered by Higgins [**14**]).

We can also exhibit an infinite family of polynomials that pick the prime values of f, spaced two apart, and generate forty consecutive, distinct primes. Consider the polynomial, for all $n \in \mathbb{Z}$, that comes from $f(x)$ via the translation $x \to 2x - 39 - 2n$:

$$h_n(x) = 4x^2 - (8n + 154)x + 4n^2 + 154n + 1523.$$

This has discriminant $-2^2 \cdot 163$, and produces forty consecutive, distinct primes for the values:

$$h_n(n+x) = \begin{cases} f(38-2x) & \text{for } x = 0,1,\ldots,19, \\ f(2x-39) & \text{for } x = 20,\ldots,39. \end{cases}$$

These polynomials just pick the prime values of $f(x)$, spaced two apart. In particular, if $n = 0$, then $h_0(x) = 4x^2 - 154x + 1523$ produces forty consecutive, distinct primes: $h_0(0) = 1523 = f(38)$, $h_0(1) = 1373 = f(36), \ldots,$ $h_0(19) = 41 = f(0)$, $h_0(20) = 43 = f(1)$, $h_0(21) = 53 = f(3), \ldots,$ $h_0(39) = 1601 = f(39)$.

Similarly, we can show that there is an infinite family of polynomials that generate eighty consecutive primes (each repeated twice) from the consecutive, prime values of f. Consider the following polynomial, for all $n \in \mathbb{Z}$, achieved from $f(x)$ via $x \to x - 40 - n$, and having discriminant -163:

$$k_n(x) = x^2 - (2n + 79)x + n^2 + 79n + 1601$$

is prime for $k_n(n+x) = k_n(n + 79 - x) = f(39 - x)$ for all $x = 0, 1, 2, \ldots, 39$. Thus, k_n picks the values of $f(39) = 1601$ to $f(0) = 41$ in descending order, then repeats the values from 41 to 1601 in ascending order. For instance, if $n = 0$, then we get the polynomial discovered by Escott in 1899 [**8**]:

$$k_0(x) = x^2 - 79x + 1601,$$

which produces eighty consecutive primes for the values $x = 0, 1, 2, \ldots, 79$. Also, if we let $n = 1460$, then we get the polynomial found by Miot in 1912 (see [**7**, p. 421]):

$$k_{1460}(x) = x^2 - 2999x + 2248541$$

is prime for the eighty consecutive values: $x = 1460, 1461, \ldots, 1539$.

However, in each of these cases, the *number* of consecutive, distinct primes, forty of them, remains the same, and this is our major concern here, *not* the actual output primes themselves, which are irrelevant for our purposes.

With the preceding examples as motivation, we consider repetitions of output prime values for a string of input values of a *given* polynomial to be *cheating*. By this we mean that the *real* test for prime-production comes from a quadratic polynomial's ability to create a string of *distinct*, consecutive prime output values, *irrespective of what those actual prime values happen to be*. Therefore, with Euler's polynomial as the template, we study *quadratics* $f(x) = ax^2 + bx + c$, with discriminant $\Delta = b^2 - 4ac < 0$. These polynomials produce distinct primes for a string of

values starting with $x = 0$, which we call an *initial string of distinct prime values*. Therefore, we assume that $c > 0$. Also, motivated by $g_n(x)$, $h_n(x)$, and $k_n(x)$, we now formalize the notion of the maximum number of distinct prime values in an initial string produced by a given quadratic.

Definition 2.1. *Consider* $F(x) = ax^2 + bx + c \quad (a, b, c \in \mathbb{Z})$, $a \neq 0$, *and suppose* $|F(x)|$ *is prime for all integers* $x = 0, 1, \ldots, \ell - 1$. *If* $\ell \in \mathbb{N}$ *is the* smallest *value such that* $|F(\ell)|$ *is composite,* $|F(\ell)| = 1$, *or* $|F(\ell)| = |F(x)|$ *for some* $x = 0, 1, \ldots, \ell - 1$, *then* $F(x)$ *is said to have* prime-production length ℓ.

For instance, the prime-production length for the polynomials $k_0(x)$, $g_0(x)$, and $f(x) = k_{-40}(x)$ is 40.

Remark 2.1. In this section, we are not concerned about the absolute values in Definition 2.1 since we deal only with positive-valued polynomials $F_\Delta(x)$. Also, $F_\Delta(x) = 1$ if and only if $x = 0$ and $A = 1$, namely $\Delta = -3$. However, in the next section, we will have to consider the possibility that the polynomials may be negative or 1, since we deal there with some positive discriminants.

From the perspective of prime-production length, it suffices to look at discriminants $\Delta \equiv 1 \pmod 4$, since the other case is trivial. To see this, assume that $\Delta < -4$, where $\Delta \equiv 0 \pmod 4$ is the discriminant of $F(x) = ax^2 + bx + c$. If c is even, then $F(2)$ is even and composite (since $F(0) = c$ must be 2, given that we are assuming $\ell \geq 1$). If c is odd, then $F(1)$ is even and composite (since b must be even when $\Delta \equiv 0 \pmod 4$). Hence, prime-production length does not exceed 2 when $\Delta \equiv 0 \pmod 4$, $\Delta < -4$.

We begin with an investigation of monic polynomials. With Euler's polynomial as the *template*, we concentrate upon the Eulerian form

$$F_\Delta(x) = x^2 + x + A,$$

where $\Delta = 1 - 4A$. We observe that $F_\Delta(A - 1) = A^2$. Therefore, the prime-production length for $F_\Delta(x)$ can be at most $A - 1$. In general, we have:

Proposition 2.1. *If* $\ell \in \mathbb{N}$ *is the prime-production length of* $F_\Delta(x)$, *then* $\ell \leq A - 1$, *where* $\Delta = 1 - 4A$. *If* p *is the least odd prime such that* $\Delta \equiv x^2 \pmod p$ *for some* $x \in \mathbb{Z}$, *then* $\ell < p$. *Furthermore, if* $\Delta \neq -7$, *and* $\ell \geq (A - 1)/2$, *then* $A = p$.

Proof: The first statement follows from the discussion preceding the statement of the proposition. Since $\Delta \equiv x^2 \pmod p$ for some $x \in \mathbb{Z}$, then we may assume that $0 \leq x < p$, without loss of generality. If $x = 0$, then $p \mid \Delta$. Thus, $F_\Delta((p - 1)/2) = (p^2 - \Delta)/4 \equiv 0 \pmod p$. If $\ell > (p - 1)/2$, then $(p^2 - \Delta)/4 = p$, that is $\Delta = p^2 - 4p < 0$. Therefore, $\Delta = -3 = -p$, contradicting that $\ell > (p - 1)/2$. Hence, $\ell \leq (p - 1)/2$. If $\ell \geq (A - 1)/2$, and $A \neq 2$, then by the minimality of p, $F_\Delta(0) = A > p$, so $\ell > (p - 1)/2$, a contradiction. We may now assume that $x > 0$.

By replacing x by $p - x$ if necessary, we may assume that $x < p$ is odd. If $x = 2n + 1$ for some integer $n \geq 0$, then $\Delta \equiv (2n + 1)^2 \pmod p$. Therefore, $F_\Delta(n) = n^2 + n + (1 - \Delta)/4 = ((2n + 1)^2 - \Delta)/4 \equiv 0 \pmod p$. Furthermore,

$F_\Delta(p-1-n) \equiv F_\Delta(n) \equiv 0 \pmod{p}$. However, since $0 < x < p$, then $0 \le n < (p-1)/2$, so $F_\Delta(n) < F_\Delta(p-1-n)$. If $p-1-n < \ell$, then $F_\Delta(p-1-n) = p = F_\Delta(n)$, a contradiction. Thus, $p-1 \ge p-1-n \ge \ell$. We have established the second statement of the theorem.

If $\ell \ge (A-1)/2$, then $0 \le n < (p-1)/2 \le (A-1)/2 \le \ell$ (observing that $\Delta \equiv 1 \pmod{A}$, and that we are assuming $\ell \ge 1$, so A is prime). Thus, $F_\Delta(n) = p$. However, $F_\Delta(n) \ge F_\Delta(0) = A$. Therefore, either $A = 2$ or $A = p$. If $(1-\Delta)/4 = A = 2$, then $\Delta = -7$. ∎

With this setup and the Euler polynomial as motivation, we may now state our first goal:

Goal #1: Classify all polynomials of the form $x^2 + x + A$ $(x \in \mathbb{Z})$ that have prime-production length $\ell = A - 1$.

By Proposition 2.1, we know that, if $\ell = A - 1$, then A is the least odd prime for which $\Delta < -7$ is a quadratic residue. For example, $A = 41$ is the least such prime for $\Delta = -163$.

Before going further, we wish to instill in the reader some further appreciation of the polynomial $F_\Delta(x) = x^2 + x + A$, and our quest to achieve Goal #1. We may ask: What is the *largest* number of consecutive prime values that polynomials of the form $F_\Delta(x)$ can assume? The following gives evidence that the answer is surprising: *Any* number of consecutive values may be assumed. To do this, we need to understand something called the "prime k-tuples conjecture". This is a generalization of the "twin primes conjecture" which says that p and $p+2$ are both prime infinitely often. One might ask: Can we have p, $p+2$, and $p+4$ simultaneously prime, infinitely often? The answer is clearly "No", since one of them must be a multiple of 3. A similar argument proves that, in the sequence p, $p+2$, $p+6$, $p+8$, $p+24$, one of the values is always divisible by 5. Thus, we must look further for a generalization of the twin primes conjecture, since it is not so straightforward.

Let $R = \{r_1, \dots, r_k\}$ with $r_i \in \mathbb{Z}$ for $i = 1, 2, \dots, k$. If q is a prime such that $\prod_{i=1}^k (n + r_i) \equiv 0 \pmod{q}$, for each $n \in \mathbb{Z}$ with $1 \le n \le q$, then there cannot exist infinitely many primes p such that $\{p + r_i\}_{i=1}^k$ are all simultaneously prime. If such a prime q exists, then we call R *inadmissible*, and otherwise we call R *admissible*. Another way of looking at this is that R is admissible if and only if, for all primes q, there exists an integer a_q with $1 \le a_q \le q$, such that $\prod_{i=1}^k (a_q + r_i) \not\equiv 0 \pmod{q}$. One might say that, if there is no good reason why $p + r_1, p + r_2, \dots, p + r_k$ cannot all be prime infinitely often, then they *should be*. More precisely, we have the following.

Conjecture 2.1 (The Prime k-tuples Conjecture). *If R is an admissible set, then there are infinitely many integers n such that $n + r$ is prime for each $r \in R$. (The twin prime conjecture is the case $R = \{0, 2\}$).*

Remark 2.2. Dickson looked at questions concerning this conjecture as early as 1904 [**7**, p. 417]. In 1923, a landmark paper of Hardy and Littlewood [**12**] appeared, in which they introduced a function based upon admissible sets, which Hensley and Richards [**13**] were able to use a half century later. Hensley and Richards showed

that there is a conflict between the prime k-tuples conjecture and what we call Conjecture A: $\pi(x+y) \leq \pi(x) + \pi(y)$ where $x,y \in \mathbb{Z}$, $x, y \geq 2$, and $\pi(x)$ denotes the number of primes not exceeding x. At least one of these (most likely Conjecture A) must be false.

We are now in a position to provide evidence that there can be arbitrarily many consecutive primes in an initial range that can be assumed by polynomials of the form $x^2 + x + A$. In other words, the prime-production length of polynomials of the form $F_\Delta(x)$ is unbounded.

Theorem 2.1.[2] *If Conjecture* 2.1 *holds, then for any positive integer B, there exists a quadratic polynomial of the form $F_\Delta(x) = x^2 + x + A$, such that $F_\Delta(x)$ is prime for all integers x with* $0 \leq x \leq B$.

Proof: Let $r_j = j^2 + j$ for $j = 0, 1, 2, \dots, B$.
Claim. The set $\{r_j\}_{j=0}^{B}$ is admissible.

If $q = 2$, then let $a_q = 1$. Since each r_j is even, then $\prod_{j=0}^{B}(r_j + 1)$ is odd. For each odd prime q, let $b_q \equiv 1 \pmod 4$ be a quadratic nonresidue modulo q (which the Chinese remainder theorem allows us to select) and set $a_q = (1 - b_q)/4$. If $\prod_{j=0}^{B}(r_j + a_q) \equiv 0 \pmod q$, then for some j with $0 \leq j \leq B$, $r_j + a_q \equiv 0 \pmod q$. In other words, $r_j \equiv -a_q \pmod q$. Therefore, $(2j+1)^2 = 4r_j + 1 \equiv 1 - 4a_q = b_q \pmod q$, a contradiction. This establishes the Claim.

Conjecture 2.1 ensures that there exist arbitrarily large values of A for which $\{r_j + A\}_{j=0}^{B}$ are primes. For such an A, $F_\Delta(x) = x^2 + x + A$ is prime for $x = 0, 1, 2, \dots, B$. ∎

Theorem 2.1 gives us the (theoretical) hope of finding a quadratic polynomial of the form $F_\Delta(x) = x^2 + x + A$ that generates not only more primes than Euler's polynomial, but also as many as we like! Yet on the practical side, no polynomial of the form $F_\Delta(x)$ with prime-production length more than forty has yet been found. Recent efforts by Lukes, Patterson, and Williams **[18]** have shown that, if such an $F_\Delta(x)$ exists, then $A > 10^{18}$. This shows the dichotomy between theory and practice, in this regard. Furthermore, Proposition 2.1 tells us that the length of prime-production for $F_\Delta(x)$ is bounded by $A - 1$. In view of Theorem 2.1, the reader may think that this is a contradiction since Theorem 2.1 seems to tell us that the prime-production length is unbounded. However, the point is simply that the B in Theorem 2.1 must be less than $A - 1$. Therefore, although the prime-production length for a *fixed* polynomial is bounded by $A - 1$, the number of primes that can be taken on by such polynomials is unbounded. For instance, the work in **[18]** shows that if we want a prime-production length of $B = 41$, then the

[2] Theorem 2.1 was communicated to this author in a letter from Andrew Granville dated February 2, 1989. He gave permission for it to be included in Louboutin, Mollin and Williams' work **[17]**, where the reader will find more details, and motivation arising from a search for prime-producing quadratic polynomials, when $\Delta > 0$. Also, A. Balog credits Granville with discovering a corollary of his main result in **[1]**. This corollary says that (unconditionally) there exist infinitely many polynomials of degree k having prime values at $2k + 1$ consecutive integers.

value of A is bigger than 10^{18}. All that Proposition 2.1 says is that $B = 41$ must be less than $A - 1$. Since $A > 10^{18}$, this is no barrier to finding such a polynomial.

We now focus our attention upon finding those $F_\Delta(x)$ that are prime for $x = 0, 1, \ldots, A - 2$. Questions concerning prime-producing quadratic polynomials become interesting only if we look at irreducible polynomials. (Recall that an irreducible polynomial over $\mathbb{Q}$ is a polynomial $f(x)$ for which there are no factorizations $f(x) = g(x)h(x)$ into polynomials $g(x)$ and $h(x)$ of positive degree with coefficients in the rational field $\mathbb{Q}$.) Even more, we should also assume that there is no prime p that divides $F_\Delta(x)$ for all $x \in \mathbb{Z}$ since, for example, $x^2 + x + 4$ is irreducible but *always even*.[3]

The criterion that allows us to achieve our Goal #1 is called the *Rabinowitsch criterion*, proved in 1913. To understand what the Rabinowitsch criterion says, we must now introduce the notion of a *class number*. To do this, we must first understand what a ring of integers is, since the class number is a measure of unique factorization therein. Independent of polynomials, we may define a *discriminant* as follows. If $D \neq 1$ is a square-free integer, and

$$\Delta = \begin{cases} 4D & \text{if } D \not\equiv 1 \pmod 4, \\ D & \text{otherwise}, \end{cases}$$

then Δ is called a *fundamental discriminant* or *field discriminant*, since we may form $K = \mathbb{Q}(\sqrt{D})$ from the adjunction of a root of the irreducible polynomial $x^2 - D$ to $\mathbb{Q}$. This K is called a *quadratic field with discriminant* Δ and *associated radicand* D; K is called a *real quadratic field* when $\Delta > 0$, and K is called a *complex quadratic field* when $\Delta < 0$. A complex number is an *algebraic integer* if it is the root of a monic polynomial with coefficients in $\mathbb{Z}$ (which we call the *rational integers* to distinguish them from higher order algebraic integers). If f is a monic polynomial over $\mathbb{Z}$ of least degree having an algebraic integer α as a root, then f is irreducible over $\mathbb{Q}$. This unique polynomial is called the *minimal polynomial* of α over $\mathbb{Q}$, and so the previous statement is equivalent to saying that the minimal polynomial of α over $\mathbb{Q}$ has coefficients in $\mathbb{Z}$. Moreover, the set of all algebraic integers in the complex field $\mathbb{C}$ is a *ring*, which we denote by $\mathbb{A}$. Finally, $\mathbb{A} \cap K = \mathcal{O}_\Delta$ defines the *ring of integers* of the quadratic field K of discriminant Δ.

Now we introduce the basic notion of an ideal. A subset I of $\mathcal{O}_\Delta$ is an *ideal* therein if I is closed under addition and also multiplication from $\mathcal{O}_\Delta$. In other words, $\alpha \pm \beta \in I$ whenever $\alpha, \beta \in I$, and $\alpha\gamma \in I$ whenever $\alpha \in I$ and $\gamma \in \mathcal{O}_\Delta$.

We say that two $\mathcal{O}_\Delta$-ideals I and J are *equivalent* if there exist nonzero $\alpha, \beta \in \mathcal{O}_\Delta$ such that $(\alpha)I = (\beta)J$. This equivalence relation partitions the $\mathcal{O}_\Delta$-ideals I into disjoint ideal classes $\{I\}$, which form a finite abelian group $\mathcal{C}_\Delta$, called the *class group* of $\mathcal{O}_\Delta$ (or simply of K). The *order* (cardinality) of $\mathcal{C}_\Delta$, denoted h_Δ, is called the *class number* of K.

Now we are ready to state the Rabinowitsch criterion, which tells us when $h_\Delta = 1$ for $\Delta < 0$. In other words, it tells us how the prime-producing capacity of $F_\Delta(x)$ for $\Delta < 0$ is intimately linked to the solution of Gauss's class number one problem for complex quadratic fields. We use $\lfloor x \rfloor$ to denote the greatest integer less than or equal to x, namely the "floor" of x.

[3]In [4], Bouniakowsky conjectured that, if a polynomial $p(x) \in \mathbb{Z}[x]$ is irreducible and $N = \gcd(\{p(x) \colon x \in \mathbb{Z}\})$, then $p(x)/N$ takes on prime values for infinitely many $x \in \mathbb{Z}$.

Theorem 2.2 (Rabinowitsch's Criterion [23]).[4] *Let* $\Delta < 0$ *be a discriminant with* $\Delta \equiv 1 \pmod 4$. *Then* $F_\Delta(x) = x^2 + x + (1 - \Delta)/4$ *is prime for all* $x \in \mathbb{Z}$ *with* $0 \leq x \leq \lfloor |\Delta|/4 - 1 \rfloor$ *if and only if* $h_\Delta = 1$. (*We call* $\lfloor |\Delta|/4 - 1 \rfloor$ *the* Rabinowitsch bound.)

We see that Euler's polynomial $k_{-40}(x) = x^2 + x + 41$ fits quite nicely into the criterion. In point of fact, it is the last one to do so, as the following solution of the class number one problem for complex quadratic fields shows. This was solved independently by Baker and Stark, anticipated by Heegner (see [**19**, Chapter 4, pp. 105–128]).

Theorem 2.3 (Gauss's Class Number One Problem for $\Delta < 0$). $h_\Delta = 1$ *if and only if* $-\Delta \in \{3, 4, 7, 8, 11, 19, 43, 67, 163\}$.

Proof: See [**6**, Theorem 7.30, p. 149]. ■

Remark 2.3. It is not very well-known that Gauss's class number one problem was solved by Landau in 1903. Although Theorem 2.3 is known as the solution of Gauss's class number one problem, this is not entirely accurate. The problem is essentially one of interpretation. Gauss's discriminant as defined in footnote 1 makes $\Delta = b^2 - ac$ even. Thus, Landau proved that $h_\Delta > 1$ for $\Delta = b^2 - ac < -7$, a much simpler problem than the one solved in Theorem 2.3.

Theorems 2.2–2.3 show us that $F_\Delta(x) = x^2 + x + A$ cannot be consecutively prime for $x = 0, 1, 2, \dots, A - 2$ when $A > 41$. The Euler polynomial tops the list with a prime-production length of 40. Thus, although Theorem 2.1 tells us that, given any integer $B > 0$, there exists a polynomial $x^2 + x + A$ that is prime for all nonnegative integers $x \leq B$, Theorems 2.2–2.3. establish that *B cannot be $A - 2$ unless* $A \leq 41$. This points to the special nature of Euler's polynomial as the *optimal* such prime-producer. We have therefore achieved Goal #1.

Remark 2.4. The reader should understand that, one day, we may discover a polynomial $F_\Delta(x)$ with $A > 0$ having prime-production length bigger than 40, probably *much* bigger, if we believe Theorem 2.1. However, A will be *enormous* relative to the prime-production length. The search for even one such polynomial with prime-production length 41 has already shown, via the work in [**18**], that A must be bigger than 10^{18}.

Now we turn to the non-monic case for prime-producing quadratics. Legendre knew that $2x^2 + 29$ is prime for $x = 0, 1, \dots, 28$ [**7**, p. 420]. Also, Lévy [**16**]

[4] A two-line proof of this fact for a more general case is in [**19**, Theorem 4.1.2, p. 108]. An interesting anecdote about Rabinowitsch comes from Mordell [**21**]: "In 1923 I attended a meeting of the American Mathematical Society held at Vassar College in New York State. Someone called Rainich from the University of Michigan at Ann Arbor, gave a talk upon the class number of quadratic fields, a subject in which I was very much interested. I noticed that he made no reference to a rather pretty paper written by Rabinowitz from Odessa and published in Crelle's journal. I commented upon this. He blushed and stammered and said, "I am Rabinowitz". He had moved to the U.S.A. and changed his name...". Thanks go to Alf van der Poorten for making me aware of Mordell's paper. The spelling of Rabinowitsch used in this article coincides with that in Crelle [**23**].

discovered in 1914 that $3x^2 + 3x + 23$ is prime for $x = 0, 1, \ldots, 21$. More recently, Van der Pol and Speziali [26] observed in 1951 that $6x^2 + 6x + 31$ is prime for $x = 0, 1, \ldots, 28$. Why do these non-monic polynomials generate these initial strings of primes? Are there more of them with even larger prime-production lengths? To answer these questions, we first observe that these polynomials share a common shape and prime-production length: $2x^2 + 29 = qx^2 - \Delta/(4q)$, $\ell = \lfloor |\Delta|/(4q) \rfloor = 29$, where $q = 2$ and $\Delta = -232$; $3x^2 + 3x + 23 = qx^2 + qx + (q^2 - \Delta)/(4q)$, $\ell = \lfloor |\Delta|/(4q) \rfloor = 22$, where $q = 3$ and $\Delta = -267$; and $6x^2 + 6x + 31 = qx^2 + qx + (q^2 - \Delta)/(4q)$, $\ell = \lfloor |\Delta|/(4q) \rfloor = 29$, where $q = 6$ and $\Delta = -708$. This motivates the following.

Definition 2.2. *If Δ is a fundamental discriminant, and $q \geq 1$ is a square-free divisor of Δ, then we call*

$$F_{\Delta,q}(x) = \begin{cases} qx^2 - \Delta/(4q) & \text{if } 4q|\Delta, \\ qx^2 + qx + (q^2 - \Delta)/(4q) & \text{otherwise}, \end{cases}$$

the qth Euler-Rabinowitsch polynomial. (*Note that, for $q = 1$, and Δ odd, $F_{\Delta,1}(x)$ is the Eulerian form that began our discussion.*)

We also let $F(\Delta, q)$ denote the maximum number of (not necessarily distinct) primes dividing $F_{\Delta,q}(x)$ for any rational integer x with $0 \leq x \leq \lfloor |\Delta|/(4q) - 1 \rfloor$. We call $\lfloor |\Delta|/(4q) - 1 \rfloor$, the qth Rabinowitsch bound. (*Note that, for $q = 1$, this is the Rabinowitsch bound in Theorem* 2.2.)

The three non-monic prime-producers, found by Legendre, Lévy, and Van der Pol and Speziali have the same shape, and they each have prime-production lengths $\ell = \lfloor |\Delta|/(4q) \rfloor$. These three quadratics also share another feature: $2x^2 + 29$ has discriminant $-232 = -2^3 \cdot 29$ ($n = 2$ distinct prime divisors) and $h_{-232} = 2 = 2^{n-1}$; $3x^2 + 3x + 23$ has discriminant $-267 = -3 \cdot 89$ ($n = 2$ distinct prime divisors) and $h_{-267} = 2 = 2^{n-1}$; and $6x^2 + 6x + 31$ has discriminant $-708 = 2^2 \cdot 3 \cdot 59$ ($n = 3$ distinct prime divisors) and $h_{-708} = 2^2 = 2^{n-1}$. This motivates the statement of our second goal.

Goal #2: Let $\Delta < 0$ be a fundamental discriminant having $N + 1$ distinct prime factors, and let q be the product of all of them, excluding the largest one. Find all polynomials $F_{\Delta,q}(x)$ such that $F(\Delta, q) = 1$ and $h_\Delta = 2^N$. In other words, find all $F_{\Delta,q}(x)$ that are prime for $x = 0, 1, \ldots, \lfloor |\Delta|/(4q) - 1 \rfloor$ with $h_\Delta = 2^N$.

Goal #2 seeks those $F_{\Delta,q}(x)$ with prime-production length $\ell \geq \lfloor |\Delta|/(4q) \rfloor$ and $h_\Delta = 2^N$. If $q = 1$, then the solution to Goal #2 for $\Delta \equiv 1 \pmod 4$ is the solution to Goal #1, since the only values for which $\ell = \lfloor |\Delta|/4 \rfloor = A - 1 \geq 1$, and $h_\Delta = 1$ are those $\Delta \equiv 1 \pmod 4$ ($\Delta < -3$) in Theorem 2.3. If $q = 1$ and $\Delta \equiv 0 \pmod 4$, then $\Delta = -8$, for which $\ell = 2$, is the only possibility for $\ell \geq \lfloor |\Delta|/4 \rfloor \geq 1$, given the discussion preceding Proposition 2.1. Therefore, we concentrate upon the (non-monic) case where $q > 1$. Not surprisingly, the achievement of Goal #2 requires a generalization of Rabinowitsch's criterion. In order to do this, we must go back to Gauss, whose classical result we state in modern terminology.

Theorem 2.4. *Suppose $\Delta < 0$ is a fundamental discriminant and Δ has $N + 1$ distinct prime factors. Then $h_\Delta = 2^N$ if and only if every element of $\mathscr{C}_\Delta$ has order* 1 *or* 2.

Proof: See [19, Theorem 1.3.3, p. 16]. ∎

When every element of $\mathscr{C}_\Delta$ has order 2, we say that it has *exponent* 2, denoted $e_\Delta = 2$. In general, the *exponent* is the smallest positive integer e_Δ such that $\{I\}^{e_\Delta} = 1$ for all $\{I\} \in \mathscr{C}_\Delta$, where 1 denotes the trivial ideal class $\{\mathscr{O}_\Delta\}$. If every element has order 1, then $h_\Delta = e_\Delta = 1$, so $\mathscr{C}_\Delta = 1$. Therefore, we may restate Gauss's result as follows:

Theorem 2.5 (Theorem 2.4 restated: **Gauss's Exponent 2 Theorem**). *Suppose $\Delta < 0$ is a fundamental discriminant and Δ has $N + 1$ distinct prime factors. Then $h_\Delta = 2^N$ if and only if $e_\Delta \leq 2$.*

Remark 2.5. For those readers who prefer the theory of binary quadratic forms, Gauss's result says that, when $\Delta < 0$, there is one class per genus in the *form* class group of discriminant Δ if and only if this *form* class group has exponent less than or equal to 2. Euler was interested in such discriminants, which he called *convenient numbers* or *numeri idonei*. Gauss listed sixty-five such numbers in [**11**, §303]. Today we know that this list is complete under the assumption of the generalized Riemann hypothesis, see [**19**, pp. 172–186]. Finally, for those who prefer class-field theory, an *ideal* class group has $e_\Delta \leq 2$ if and only if the Hilbert class field is equal to the genus field [**6**, p. 127 ff].

Is the Rabinowitsch criterion merely an isolated curiosity? Is there some deeper underlying phenomenon? To answer these questions, we developed the following result, which not only generalizes Theorem 2.2, but also incorporates Theorem 2.5. The values of $\Delta = -3, -4$ are excluded in the following for trivial technical reasons.

Theorem 2.6 (Mollin [20]). *Let $\Delta < -4$ be a fundamental discriminant that has $N + 1$ distinct prime factors, with p being the largest, and suppose $q \geq 1$ is a square-free divisor of Δ, having $M \geq 0$ distinct prime factors all less than p. Then $e_\Delta \leq 2$ if and only if $F(\Delta, q) = N + 1 - M$ and $h_\Delta = 2^{F(\Delta, q)+M-1}$.*

Thus, the Rabinowitsch criterion, an application of Theorem 2.6 for $M = 0$, may be restated as follows.

Corollary 2.1. *Let $\Delta < -4$ be a fundamental discriminant. Then $h_\Delta = 1$ if and only if $F(\Delta, 1) = 1$.*

We also get the more recent class number 2 phenomenon by R. Sasaki (see [**19**, p. 112]) another application of Theorem 2.6 for $M = 0$.

Corollary 2.2. *Let $\Delta < 0$ be a fundamental discriminant. Then $h_\Delta = 2$ if and only if $F(\Delta, 1) = 2$.*

Corollaries 2.1–2.2 are actually very special because, if $e_\Delta \leq 2$, then $h_\Delta \leq 2$ if and only if $F(\Delta, 1) = h_\Delta$. Therefore, the fields for which $\Delta < 0$ and $h_\Delta \leq 2$ are uniquely characterized by these corollaries.

Remark 2.6. In the language of binary quadratic forms, the conclusion of Theorem 2.6 says: There is one class per genus in the ideal class group of discriminant Δ if and only if $F(\Delta, q) = N + 1 - M$ and $h_\Delta = 2^{F(\Delta, q)+M-1}$.

Theorem 2.5 tells us that $e_\Delta \le 2$ if and only if $h_\Delta = 2^N$. Thus, one direction in Theorem 2.6 is clear. Let q be the product of all prime factors of Δ *excluding* the largest (namely, when $M = N$). What is new and revealing is that $h_\Delta = 2^N$ implies $F(\Delta, q) = 1$. In particular, we have proved the striking result that $F(\Delta, q) = 1$ when $e_\Delta \le 2$, that is $F_{\Delta,q}(x)$ is always prime up to the qth Rabinowitsch bound. This links $e_\Delta \le 2$ to the prime-production length of $F_{\Delta,q}(x)$. Thus, Theorem 2.6 shows that the Rabinowitsch criterion underlies a deeper phenomenon: a criterion for $e_\Delta \le 2$ in terms of the factorization of $F_{\Delta,q}(x)$. This is enough information to achieve Goal #2. Theorem 2.3 identifies those $\Delta < 0$ for which $e_\Delta = h_\Delta = 1$. From the work of Weinberger **[27]**, we know an upper bound on the number of discriminants $\Delta < 0$ for which $e_\Delta = 2$, under the assumption of the generalized Riemann hypothesis. Therefore, a computer search can provide us with a complete list of them.

Tables: The tables are broken down into congruences modulo 4 of the radicand D. In each of them, q is the product of all distinct primes dividing the discriminant Δ, *excluding* the largest prime p. Moreover, ℓ is the prime-production length of $F_{\Delta,q}(x)$, and h_Δ is the class number. Finally, concerning Remark 2.5, the reader will notice that the number of discriminants in Theorem 2.3 and Tables 1–3 is sixty-five. Thirty-five of these values are on Gauss's list of *convenient* numbers. These are the fundamental discriminants Δ for which either $\Delta \equiv 0 \pmod 4$ or $\Delta = -15$. The balance of the values in Gauss's list are discriminants arising from a more general situation (see **[19**, pp. 112–128]). Finally, the reader should note that $\ell = \lfloor |\Delta|/(4q) \rfloor$, in both Theorem 2.3 and in Tables 1–2, and $\lfloor |\Delta|/(4q) \rfloor + 3 \ge \ell \ge \lfloor |\Delta|/(4q) \rfloor$ in Table 3.

In Table 1, $6x^2 + 6x + 31$ is the optimal prime-producer. In Table 2, the optimal prime-producer is $2x^2 + 29$. The optimal prime-producer in Table 3 is $3x^2 + 3x + 23$. Thus, the optimal prime-producers are the polynomials that motivated our quest to achieve Goal #2! Therefore, these ubiquitous examples in the literature, which have appeared as isolated curiosities with no complete and

TABLE 1. $D \equiv 3 \pmod 4$

$\|D\|$	h_Δ	p	$F_{\Delta,q}(x)$	ℓ
5	2	5	$2x^2 + 2x + 3$	2
13	2	13	$2x^2 + 2x + 7$	6
21	4	7	$6x^2 + 6x + 5$	3
33	4	11	$6x^2 + 6x + 7$	5
37	2	37	$2x^2 + 2x + 19$	18
57	4	19	$6x^2 + 6x + 11$	9
85	4	17	$10x^2 + 10x + 11$	8
93	4	31	$6x^2 + 6x + 17$	15
105	8	7	$30x^2 + 30x + 11$	3
133	4	19	$14x^2 + 14x + 13$	9
165	8	11	$30x^2 + 30x + 13$	5
177	4	59	$6x^2 + 6x + 31$	29
253	4	23	$22x^2 + 22x + 17$	11
273	8	13	$42x^2 + 42x + 17$	6
345	8	23	$30x^2 + 30x + 19$	11
357	8	17	$42x^2 + 42x + 19$	8
385	8	11	$70x^2 + 70x + 23$	5
1365	16	13	$210x^2 + 210x + 59$	6

$F_{\Delta,q}(x)$ is prime for all nonnegative integers $x < \ell$

TABLE 2. $D \equiv 2 \pmod 4$

$\|D\|$	h_Δ	$p = \ell$	$F_{\Delta,q}(x)$
6	2	3	$2x^2 + 3$
10	2	5	$2x^2 + 5$
22	2	11	$2x^2 + 11$
30	4	5	$6x^2 + 5$
42	4	7	$6x^2 + 7$
58	2	29	$2x^2 + 29$
70	4	7	$10x^2 + 7$
78	4	13	$6x^2 + 13$
102	4	17	$6x^2 + 17$
130	4	13	$10x^2 + 13$
190	4	19	$10x^2 + 19$
210	8	7	$30x^2 + 7$
330	8	11	$30x^2 + 11$
462	8	11	$42x^2 + 11$

$F_{\Delta,q}(x)$ is prime for all nonnegative integers $x < \ell$

TABLE 3. $D \equiv 1 \pmod 4$

$\|D\|$	h_Δ	p	$F_{\Delta,q}(x)$	ℓ
15	2	5	$3x^2 + 3x + 2$	1
35	2	7	$5x^2 + 5x + 3$	2
51	2	17	$3x^2 + 3x + 5$	4
91	2	13	$7x^2 + 7x + 5$	4
115	2	23	$5x^2 + 5x + 7$	6
123	2	41	$3x^2 + 3x + 11$	10
187	2	17	$11x^2 + 11x + 7$	6
195	4	13	$15x^2 + 15x + 7$	3
235	2	47	$5x^2 + 5x + 13$	12
267	2	89	$3x^2 + 3x + 23$	22
403	2	31	$13x^2 + 13x + 11$	10
427	2	61	$7x^2 + 7x + 17$	16
435	4	29	$15x^2 + 15x + 11$	7
483	4	23	$21x^2 + 21x + 11$	6
555	4	37	$15x^2 + 15x + 13$	9
595	4	17	$35x^2 + 35x + 13$	4
627	4	19	$33x^2 + 33x + 13$	5
715	4	13	$55x^2 + 55x + 17$	6
795	4	53	$15x^2 + 15x + 17$	13
1155	8	11	$105x^2 + 105x + 29$	5
1435	4	41	$35x^2 + 35x + 19$	10
1995	8	19	$105x^2 + 105x + 31$	7
3003	8	13	$231x^2 + 231x + 61$	4
3315	8	17	$195x^2 + 195x + 53$	4

$F_{\Delta,q}(x)$ is prime for all nonnegative integers $x < \ell$

detailed explanation, are now fully understood as the optimal prime-producers with discriminants $\Delta < 0$ having $e_\Delta = 2$. Thus, Tables 1–3, via Theorem 2.6, give us the complete list and Goal #2 is achieved, assuming the validity of the generalized Riemann hypothesis, of course.

Remark 2.7. Once we leave the case of exponent 2, it is still possible to have $F(\Delta, q) = N + 1 - M$. However, when $e_\Delta > 2$, the other condition in Theorem 2.6 fails, namely $h_\Delta \neq 2^{F(\Delta,q)+M-1}$. The following example illustrates these assertions.

Let $\Delta = -9867 = -3 \cdot 11 \cdot 13 \cdot 23$ with $q = 429 = 3 \cdot 11 \cdot 13$, and $M = N = 3$. Then, $F(\Delta, q) = N + 1 - M = 1$, since $F_{\Delta,q}(x) = 429x^2 + 429x + 113$ is prime for all nonnegative integers $x \leq \lfloor |\Delta|/(4q) - 1 \rfloor = 5$. However, $h_\Delta = 2^4$ and $4 \neq F(\Delta, q) + M - 1 = 3$. Also, $e_\Delta = 4$, since $\mathscr{C}_\Delta$ is the product of two cyclic groups of order 2 and one of order 4. This example also illustrates that the two conditions in Theorem 2.6 are sharply necessary and sufficient conditions for exponent 2 to occur.

Before closing this section, it is worth revealing to the reader a criterion for $e_\Delta \leq 2$ in terms of the form of $\Delta < 0$. To do this, we must understand a little about the arithmetic of ideals in $K = \mathbb{Q}(\sqrt{D})$.

A non-unit ideal I (namely, $I \neq \mathscr{O}_\Delta$) is a *prime ideal* if, whenever I divides a product of ideals $I_1 I_2$ in $\mathscr{O}_\Delta$, then $I|I_1$ or $I|I_2$; this mimics the property of a prime $p \in \mathbb{Z}$, of course. Here "divides" means if $I|J$, then there exists an $\mathscr{O}_\Delta$-ideal H such that $J = HI$. It follows that $J \subseteq I$. Conversely, if $J \subseteq I$, then there is an ideal H such that $J = HI$. Thus, ideals that "divide" are those that contain. Therefore, I is a prime ideal if and only if it contains some non-trivial factor of any product it divides. From now on, when we say that a prime ideal $\mathscr{P}$ is *above* p, we mean that $\mathscr{P}$ divides (p). Furthermore, we call p *non-inert* if $(\Delta/p) \neq -1$, where $(*/p)$ is the *Kronecker symbol.*

Remark 2.8. Recall that the Kronecker symbol is defined as follows. If $\gcd(p, \Delta) = 1$ for a discriminant Δ, then the Kronecker symbol for $p > 2$ is just the Legendre symbol namely $(\Delta/p) = 1$ if Δ is a quadratic residue modulo p, and $(\Delta/p) = -1$ otherwise. If $p|\Delta$, then $(\Delta/p) = 0$; if $p = 2$, then $(\Delta/2) = 1$ if $\Delta \equiv 1 \pmod 8$, and $(\Delta/2) = -1$ if $\Delta \equiv 5 \pmod 8$. See [**19**, Exercise 1.1.4(b), p. 8].

If p is non-inert, then there are two cases. Either $(\Delta/p) = 1$, in which case we say that p *splits* in K, since $(p) = \mathscr{P}_1\mathscr{P}_2$ with $\mathscr{P}_1 \neq \mathscr{P}_2$; or $(\Delta/p) = 0$, in which case we say that p *ramifies* since $p|\Delta$ and $(p) = \mathscr{P}^2$ for a unique prime $\mathscr{O}_\Delta$-ideal $\mathscr{P}$ over p.

Theorem 2.7. *Suppose $\Delta < 0$ is a fundamental discriminant. Then $e_\Delta \leq 2$ if and only if, for every split prime $p < \sqrt{-\Delta/3}$, there exists a square-free divisor $q > p$ of $|\Delta|$ such that $\Delta = q^2 - 4pq$.*

Proof: [**20**, Theorem 3.1, pp. 22–23]. ∎

We close this section with an illustration of Theorem 2.7 from Table 3.

Example 2.1. *If $\Delta = -3315$, then the only split primes $p < \sqrt{-\Delta/3}$ are $p = 29, 31$. Therefore, $\Delta = 65^2 - 4 \cdot 29 \cdot 65 = 85^2 - 4 \cdot 31 \cdot 85$.*

3. DENSITY OF PRIMES. In Section 1, we were concerned with quadratic polynomials producing consecutive primes for an initial string of input values. Therein, we solved the problem for all qth Euler-Rabinowitsch polynomials, assuming the generalized Riemann hypothesis. It is natural to ask how many primes are produced by a given quadratic polynomial up to a given bound. For instance, motivated by Euler's polynomial and its generalizations, we would like to

know the number $P_A(n)$ of primes assumed by $F_\Delta(x) = x^2 + x + A$ for $x = 0, 1, 2, \ldots, n$, with $\Delta = 1 - 4A < 0$. We showed in Section 1 that $P_A(A-2) = A - 1$ if and only if $A \in \{1, 2, 3, 5, 11, 17, 41\}$, in other words for $\Delta \in \{-3, -7, -11, -19, -43, -67, -163\}$. Before it was known that $A = 41$ is the largest value for which this phenomenon occurs, Dick Lehmer, in 1936, attempted to find a larger such value of A. First he observed that A must be odd in order for $F_\Delta(x)$ to be prime, so $-\Delta \equiv 3 \pmod 8$. From the discussion leading up to Proposition 2.1, we see that $F_\Delta(x)$ can be divisible only by non-inert primes (also see [**19**, Lemmas 4.1.2–4.1.3, p. 118]). Lehmer exploited this fact by observing that, since a prime q with $(\Delta/q) = -1$ cannot divide $F_\Delta(x)$, then $F_\Delta(x)$ would be a prime a larger percentage of the time for a $\Delta < 0$ when $(\Delta/q) = -1$ for "enough" small primes q. Lehmer showed that the least positive integer $N_{109} \equiv 3 \pmod 4$ such that $(-N_{109}/q) = -1$ for all primes $q \le 109$ must satisfy $N_{109} > 5 \cdot 10^9$. For $P_A(A-2) = A - 1$, Lehmer proved that $-\Delta = 4A - 1 > N_{109}$, so $A = (1 - \Delta)/4 > 1.25 \cdot 10^9$. This told Lehmer that the existence of $A > 41$ with $P_A(A-2) = A - 1$ was highly unlikely. Thus, he thought that it might be more fruitful to search for an A with $P_A(40) = 41$. If such an A exists, we have noted in Section 1 that $A > 10^{18}$.

In 1939, Beeger published his discovery of $F_\Delta(x) = x^2 + x + A$, for $A = 19421$, 27941, 72491, and $\Delta = -77683, -111763, -289963$ [**2**]. He found these values by computing all positive integers N with $N \equiv 3 \pmod 8$ and $N < 10^6$ and such that $(-N/q) = -1$ for all odd primes $q \le 43$. The only such numbers are $N = 77683$, 111763, 289963. In 1939, Poletti [**22**] computed tables of values of $P_A(n)$ for these A and found that $P_{27941}(11000) = 4819$, $P_{72491}(11000) = 4923$ and $P_{41}(11000) = 4605$, so Beeger's polynomial is better at prime-production than Euler's polynomial. What is behind all of this is Hardy and Littlewood's "Conjecture F" from 1923 [**12**]:

Conjecture 3.1. *If* $\Delta = 1 - 4A$ *is not a perfect square, then*

$$P_A(n) \sim C(\Delta) \cdot n/\log(n),$$

where

$$C(\Delta) = \prod_{\text{primes } p \ge 3} (1 - (\Delta/p)/(p-1)).$$

Remark 3.1. Conjecture F is more general than this, but we have boiled it down to suit our purposes. In order to accommodate an error term, with which we are not concerned here, $n/\log(n)$ is replaced by $2\int_0^n dx/\log F_\Delta(x)$ (see [**19**, pp. 145–147] for more details). Also, recall that $f(n) \sim g(n)$ means $\lim_{n\to\infty} f(n)/g(n) = 1$, and $\log(n)$ means the *natural logarithm* to the base e.

For instance, we know from the work of Shanks[5] [**25**] that $C(-163) = 3.319773177471$ for $A = 41$, and $C(-289963) = 3.694708051836$ for $A = 72491$, confirming Beeger's observation. Thus, Conjecture 3.1 allows us to find examples of $F_\Delta(x)$ with a *high asymptotic density of prime values*. How $C(\Delta)$ is computed and extensions of Beeger's ideas to find new $F_\Delta(x)$ with high asymptotic density of

[5] The death of Daniel Shanks, at the age of seventy-nine on September 9, 1996, is a great loss to the world of mathematics. This author was fortunate enough to have known him and enjoyed his rapier wit, as well as the depth of his intellect. A pioneer and a giant in the world of computational number theory has left us.

primes, can be found in **[18]**. Therein, the authors report results obtained with the new number sieve MSSU, developed at the University of Manitoba. This acronym means *The Manitoba Scalable Sieve Unit* which employs *Very Large Scale Integration* (VLSI) circuits designed by C. Patterson at the University of Calgary (on whose Ph.D. committee this author sat). The raw aggregate sieving rate of the MSSU is more than 6.4 billion integers per second.

In 1990, Fung and Williams **[10]** found $C(\Delta) = 5.0870883$ for $\Delta = -531497118115723$ with $A = 132874279528931$, and $P_A(10^6) = 312975$. Compare this with $P_{41}(10^6) = 261080$. In this example, they found that $(\Delta/q) = -1$ for all odd primes $q \leq 179$. In **[18]**, the authors used the MSSU to search for all $|\Delta| < 10^{19}$ with $\Delta \equiv 5 \pmod 8$ and $(\Delta/q) = -1$ for all odd primes $q < 200$. The largest value of $C(\Delta)$ that they found is

$$C(\Delta) = 5.3384021$$

for

$$\Delta = -6849319464662435083.$$

Therefore, assuming Conjecture 3.1, $F_\Delta(x) = x^2 + x + A$ with

$$A = 1712329866165608771$$

has the largest asymptotic density of primes for any polynomial of this type to date.

Other authors have looked at less demanding problems involving prime-producing quadratics. For example, Boston and Greenwood **[3]** wanted to find a quadratic polynomial $f(x)$ that represents the most *distinct* primes for $0 \leq x \leq 99$. Of course, we have observed earlier that we can get longer strings if we allow repetitions than if we impose the restriction of distinctness. In **[3]**, it is observed that $x^2 - 69x + 1231$ represents ninety-five primes for $x = 0, \ldots, 99$, although not distinct. However, this is just the special case $k_{-5}(x)$ of our family $k_n(x) = x^2 - (2n + 79)x + n^2 + 79n + 1601$ introduced in Section 1. Moreover, we can improve upon their result slightly via $k_n(x)$. One can verify that $k_n(x)$ produces one hundred primes for $n \leq x \leq n + 104$, with the first eighty being consecutive, and twenty of the last twenty-five being primes. This provides an infinite family of polynomials that generate one hundred primes (eighty of which are consecutive) in intervals of one hundred and five.

The polynomial in **[3]** that generates the most *distinct values* is $41x^2 + 33x - 43321$, with $\Delta = 7105733$. This has prime values for ninety of the input values $0 \leq x \leq 99$. However, the most *consecutive* primes in this sequence is twenty-six, so the problem described in Section 1, which is the *real* test, is not addressed. The challenge left at the end of **[3]** (where the authors cite $C(\Delta)$ for earlier efforts of Fung and Williams but not the latest in **[18]**) is to find an example that gives forty-nine or even fifty consecutive, distinct primes. Theorem 2.1 says one should be able to do so.

In looking at class number one problems for real quadratic fields, the Fung polynomials

$$g_1(x) = 47x^2 - 2247x + 21647$$

and

$$g_2(x) = 103x^2 - 3945x + 34381,$$

and the Ruby polynomial

$$g_3(x) = 36x^2 - 810x + 2753$$

were found. It turns out that $g_1(x)$ and $g_2(x)$ have prime-production lengths $\ell = 43$, and $g_3(x)$ has prime-production length $\ell = 45$. (For $g_1(x)$, $C(\Delta) = 3.7006266$, where $\Delta = 979373$; for g_2, $C(\Delta) = 3.9727065$ where $\Delta = 1398053$; and for $g_3(x)$, $C(\Delta) = 3.9099833$, where $\Delta = 259668$, see [**18**, p. 118].) To date, these are the largest known consecutive, distinct, quadratic prime-producers, namely those with the largest prime-production lengths. The polynomials $g_i(x)$ were spinoffs from the investigations involving a search for $h_\Delta = 1$ when $\Delta > 0$, by this author and Hugh Williams (see [**19**, Chapter 4, pp. 129–145]). The discovery of g_1 and g_2 by Gilbert Fung, then a graduate student of Hugh Williams, was announced by this author in a lecture given during the Western Number Theory Conference at Las Vegas in 1988. Russell Ruby was in the audience, and he went home to check these polynomials. Later he found g_3.

ACKNOWLEDGMENTS. The author welcomes the opportunity to recognize support by NSERC Canada grant #A8484, and to thank the referees for suggestions that improved the exposition. In particular, one referee made detailed comments, which enabled this author to polish the final product.

REFERENCES

1. A. Balog, The prime k-tuplets conjecture on average, in *Analytic Number Theory* (Bruce C. Berndt, Harold G. Diamond, Heini Halberstam, and Adolf Hildebrand, eds.), Birkhäuser, Boston, Basel, Berlin (1990), 47–75.
2. N. G. W. H. Beeger, Report on some calculations of prime numbers, *Nieuw. Archief Wiskunde* 20 (1939), 48–50.
3. N. Boston and M. L. Greenwood, Quadratics representing primes, *Amer. Math. Monthly* 102 (1995), 595–599.
4. W. Bouniakowsky, Sur les diviseurs numériques invariables des fonctions rationelles entiers, *Acad. Sci. St. Pétersbourg Mém. Sci. Math. et Phy.* 6 (1857), 305–329.
5. H. Cohn, *A Second Course in Number Theory*, John Wiley and Sons Inc., New York, London (1962).
6. D.-A. Cox, *Primes of the Form $x^2 + ny^2$*, J. Wiley and Sons, New York (1989).
7. L. E. Dickson, *History of the Theory of Numbers, Vol. I*, Chelsea Publishing Co., New York (1971).
8. E. B. Escott, Réponses 1133, Formule d'Euler $x^2 + x + 41$ et formules analogues, *L'intermédiare des Math.* 6 (1899), 10–11.
9. L. Euler, Extrait d'une lettre de M. Euler le père à M. Bernoulli concernant le memoire imprimé parmi ceux de 1771, *Nouveaux mémoirs de l'Acadèmie des Sciences de Berlin* 1772 (1774), p. 381.
10. G. W. Fung and H. C. Williams, Quadratic polynomials which have a high density of prime values, *Math. Comp.* 55 (1990), 345–353.
11. C. F. Gauss, *Disquisitiones Arithmeticae* (English edition), Springer-Verlag, New York, Berlin (1986).
12. G. H. Hardy, and J. E. Littlewood, Some problems of 'Partitio Numerorum' III: On the expression of a number as a sum of primes, *Acta. Math.* 44 (1923), 1–70.
13. D. Hensley, and I. Richards, On the incompatibility of two conjectures concerning primes, *Proc. Symp. in Pure Math., Analytic Number Theory* 24 American Mathematical Society, Providence (1973), 123–127.
14. O. Higgins, Another long string of primes, *J. Rec. Math.* 14 (1982), 185.
15. A. M. Legendre, *Théorie des nombres*, Libraire Scientifique, A. Herman, Paris (1798), 69–76; second ed. (1808); third ed. (1830), 72–80.
16. A. Lévy, Sur les nombres premiers dérivés de trinomes du second degré, *Sphinx-Oedipe* 9 (1914), 6–7.
17. S. Louboutin, R. A. Mollin, and H. C. Williams, Class numbers of real quadratic fields, continued fractions, reduced ideals, prime-producing quadratic polynomials, and quadratic residue covers, *Can. J. Math.* 44 (1992), 824–842.
18. R. F. Lukes, C. D. Patterson, and H. C. Williams, Numerical sieving devices: their history and some applications, *Nieuw Archief voor Wiskunde* 13 (1995), 113–139.
19. R. A. Mollin, *Quadratics*, CRC Press, Boca Raton, New York, London, Tokyo (1995).
20. R. A. Mollin, Quadratic polynomials producing consecutive, distinct primes and class groups of complex quadratic fields, *Acta Arith.* 74 (1996), 17–30.

21. L. J. Mordell, Reminiscences of an octogenarian mathematician, *Amer. Math. Monthly* 78 (1971), 952–961.
22. L. Poletti, Au sujet de la décomposition des termes de la séries $z = x^2 + x + 146452961$, Sphinx 9 (1939), 83–85.
23. G. Rabinowitsch, Eindeutigkeit der Zerlegung in Primzahlfactoren in quadratischen Zahlkörpern, *J. Reine Angew. Math.* 142 (1913), 153–164.
24. P. Ribenboim, Euler's famous prime generating polynomial and the class number of imaginary quadratic fields, *L'Enseignment Math.* 34 (1988), 23–42.
25. D. Shanks, Calculation and applications of Epstein zeta functions, *Math. Comp.* 29 (1975), 271–287.
26. B. Van der Pol, and P. Speziali, The primes in $k(\zeta)$, *Indag. Math.* 13 (1951), 9–15.
27. P. J. Weinberger, Exponents of the class groups of complex quadratic fields, *Acta. Arith.* 22 (1973), 117–124.

Mathematics Department
University of Calgary
Calgary, Alberta
T2N 1N4 Canada
http://www.math.ucalgary.ca/~ramollin/
ramollin@math.ucalgary.ca

From the MONTHLY Fifty Years Ago...

Of the twenty most popular books, seven are in mathematics, all high school texts, and four others are bookkeeping and accounting. The most popular book of all is *Auto-Mechanics*; second is *Bookeeping and Accounting*; third is *A First Course in Algebra, Part I*; and fourth is *A First Course in Algebra, Part II*. The only college level books among the twenty most popular are *Principles of Accounting* and *Modern Electric and Gas Refrigeration*. All of the books except the last two are "self-teaching."

Of the twenty most popular high school correspondence courses, five are mathematics. The most popular course is Beginning Algebra; second is Review Arithmetic; and third, Bookkeeping. Trigonometry is ninth on the list....

Of the twenty most popular college correspondence courses, four are mathematics. The most popular course is College Algebra and Trigonometry; second, Introduction to Accounting; and third, English Composition. Differential Calculus ranks tenth; Differential Equations is twenty-eighth; and Integral Calculus, thirty-sixth. Twenty-four percent of all those enrolled in college correspondence courses are taking mathematics....

...a point which is very important to every mathematician ... is the decreasing mathematical content of the high school and college curricula during the past twenty or thirty years. Twenty-five years ago, every student was required to take two years of good, sound algebra in order to graduate; in addition, two more years of mathematics were required for an A.B. degree in many colleges. The present situation is a far cry from that, even in the most conservative institutions. Mathematics has gradually been removed from the various curricula until there is very little left that is useful or even recognizable. Many of the courses which are called mathematics are a disgrace to the name. They are designed for amusement, and anything which might be thought-provoking is carefully avoided.

Colonel W. E. Sewell, Mathematics in the Army Education Program
MONTHLY 54 (1947) 195–200

The Poincaré-Miranda Theorem

Wladyslaw Kulpa

Bernard Bolzano (1781–1848), the outstanding Czech thinker, philosopher and mathematician, proved that if a function f, continuous in a closed interval $[a, b]$ changes signs at the endpoints; $f(a) \cdot f(b) \leq 0$, then this function equals zero at some point of the interval. In 1883–1884, Henri Poincaré announced the following result without proof **[9]**, **[10]** (in Browder's translation **[3]**):

> "Let $f_1, \dots, f_n$ be n continuous functions of n variables $x_1, \dots, x_n$: the variable x_i is subjected to vary between the limits $+a_i$ and $-a_i$. Let us suppose that for $x_i = a_i$, f_i is constantly positive, and that for $x_i = -a_i$, f_i is constantly negative; I say there will exist, a system of values of x where all the f's vanish."

In 1886 Poincaré **[11]** published his argument on the homotopy invariance of the index, which is a basis for the proof. The result obtained by Poincaré has come to be known as the theorem of Miranda **[8]**, who in 1940 showed that it was equivalent to the Brouwer fixed point theorem. The Poincaré theorem was implicitly rediscovered in 1911 by Brouwer **[2]** who proved that

> "Under a continuous map of the unit cube into itself which displaces every point less than half a unit, the image has an interior point."

The Brouwer fixed point theorem for $n = 3$ was proved by him in 1909; an equivalent result was established earlier by Bohl **[2]** in 1904. It was Hadamard **[4]** who in 1910 gave (using the Kronecker index) the first proof for arbitrary n. In 1912 Brouwer gave another proof using the simplicial approximation technique, and notions of degree. A short and simple proof of the Bohl-Brouwer theorem was given in 1929 by Knaster-Kuratowski-Mazurkiewicz **[7]**; the proof was based on the lemma of Sperner **[12]**. In this paper, we also apply combinatorial methods of proof taken from the Sperner lemma, but our proof seems to be simpler because it does not require the notion of barycentric coordinates and barycentric subdivision.

Let us begin by discussing the Poincaré-Miranda Theorem. Let $k > 1$ be a given natural number and let $Z_k := \{i/k : i \in Z\}$, where Z denotes the set of integers. Let Z_k^n denote the Cartesian product of n copies of the set Z_k:

$$Z_k^n := \{z : \{1, \dots, n\} \longrightarrow Z_k \quad | \quad z \text{ is a map}\}$$

Using the Cartesian notation let $\mathbf{0} := (0, \dots, 0)$ be the neutral element and let $e_i := (0, \dots, 0, 1/k, 0, \dots, 0)$, $e_i(i) = 1/k$, be the i-th basic vector. Denote by $P(n)$ the set of permutations of the set $\{1, \dots, n\}$.

Definition 1. An ordered set $S = [z_0, \dots, z_n] \subset Z_k^n$ is said to be an *n-simplex* if there exists a permutation $\alpha \in P(n)$ such that

$$z_1 = z_0 + e_{\alpha(1)}, \quad z_2 = z_1 + e_{\alpha(2)}, \dots, \quad z_n = z_{n-1} + e_{\alpha(n)}.$$

Any subset $[z_0, \dots, z_{i-1}, z_{i+1}, \dots, z_n] \subset S$, $i = 0, \dots, n$, is said to be the $(n-1)$-*face* of the n-simplex S. A subset $C \subset Z_k^n$ of the form

$$C := C(k) = \left\{0, \frac{1}{k}, \dots, \frac{k-1}{k}, 1\right\}^n$$

is said to be a *combinatorial n-cube*. Define the *i-th combinatorial back and front faces* of C as

$$C_i^- := C_i^-(k) = \{z \in C : z(i) = 0\}, \quad C_i^+ := C_i^+(k) = \{z \in C : z(i) = 1\},$$

and the *boundary* as

$$\partial C := \bigcup \{ C_i^- \cup C_i^+ : i = 1, \dots, n\}$$

Definition 2. Let $S = [z_0, \dots, z_n] \subset Z_k^n$ be an n-simplex. Then for each point $z_i \in S$ there exists exactly one n-simplex $T = S[i]$ such that

$$S \cap T = \{z_0, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}.$$

We shall define the *i-neighbour* $S[i]$ of the simplex S (see Figure 1) as

(a) If $0 < i < n$, then $S[i] := [z_0, \dots, z_{i-1}, x_i, z_{i+1}, \dots, z_n]$, where $x_i = z_{i-1} + (z_{i+1} - z_i) = z_{i-1} + e_{\alpha(i+1)}$.
(b) If $i = 0$, then $S[0] := [z_1, \dots, z_n, x_0]$, where $x_0 = z_n + (z_1 - z_0)$.
(c) If $i = n$, then $S[n] := [x_n, z_0, \dots, z_{n-1}]$, where $x_n = z_0 + (z_{n-1} - z_n)$.

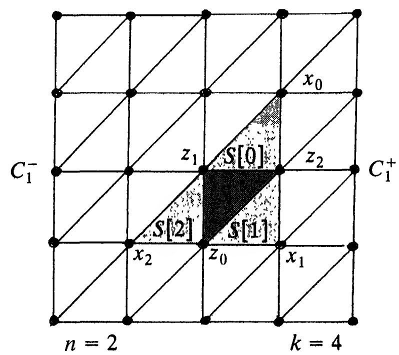


Figure 1

We leave it to the reader to prove that the n-simplexes $S[i]$ are well-defined and that they are the only possible i-neighbours of the n-simplex S.

From Definition 2 the following observation is immediate:

Observation. Any $(n-1)$-face of an n-simplex contained in the combinatorial n-cube C is an $(n-1)$-face of exactly one or two n-simplexes from C, depending on whether or not it lies on the boundary ∂C.

Let $I^n := [0,1]^n$ be the n-dimensional cube of the Euclidean space R^n and let ∂I^n be its boundary. For each $i \le n$ let us denote

$$I_i^- := \{x \in I^n: \quad x(i) = 0\}, \quad I_i^+ := \{x \in I^n: \quad x(i) = 1\},$$

the i-th opposite *faces*.

Poincaré-Miranda Theorem. *Let f: $I^n \longrightarrow R^n$, $f = (f_1, \dots, f_n)$, be a continuous map such that for each $i \le n$, $f_i(I_i^-) \subset (-\infty, 0]$ and $f_i(I_i^+) \subset [0, +\infty)$. Then there exists a point $c \in I^n$ such that $f(c) = \mathbf{0}$.*

Proof: For each $i = 1, \dots, n$ define $H_i^- := f_i^{-1}(-\infty, 0]$ and $H_i^+ := f_i^{-1}[0, \infty)$. Since for each sequence of n-simplexes $S_k \subset C(k)$, *diameter* $S_k \longrightarrow 0$ as $k \longrightarrow \infty$, in order to prove the theorem it suffices to show that for each k there exists an n-simplex $S_k \subset C(k)$ such that

$$H_i^- \cap S_k \neq \emptyset \neq H_i^+ \cap S_k \quad \text{for each } i = 1, \dots, n. \tag{1}$$

Indeed, using a compactness argument we infer that the intersection

$$H := \bigcap \{H_i^- \cap H_i^+: \quad i = 1, \dots, n\}$$

is not empty. It is clear that $f(c) = \mathbf{0}$ for each $c \in H$.

Define a map φ: $I^n \longrightarrow \{0, \dots, n\}$ by

$$\varphi(x) := max \left\{ j: x \in \bigcap_{i=0}^{j} F_i^+ \right\}, \tag{2}$$

where $F_0^+ = I^n$ and $F_i^+ = H_i^+ \setminus I_i^-$ for each $i = 1, \dots, n$. Since $I^\varepsilon \subset H_i^\varepsilon$, where $\varepsilon = +$ or $-$, the map φ has the following properties:

$$\text{if } x \in I_i^-, \text{ then } \varphi(x) < i, \text{ and if } x \in I_i^+, \text{ then } \varphi(x) \neq i - 1. \tag{3}$$

From (3) it follows that for each subset $S \subset I^n$

$$\varphi(S \cap I_i^\varepsilon) = \{0, \dots, n-1\} \text{ implies that } i = n \text{ and } \varepsilon = - . \tag{4}$$

Observe that (2) and the fact that $I^n = H_i^- \cup H_i^+$ imply that

$$\text{if } \varphi(x) = i - 1 \text{ and } \varphi(y) = i, \text{ then } x \in H_i^- \text{ and } y \in H_i^+. \tag{5}$$

Let us call a finite subset S of $l + 1$ points in the combinatorial cube $C = C(k)$ *admissible* if $\varphi(S) = \{0, \dots, l\}$. From (1) and (5) it follows that the theorem will be proved if we show that for each k there exists an admissible n-simplex $S \subset C(k)$. We shall actually prove that for each k the number $\alpha(C(k))$ of admissible n-simplexes is odd.

Our proof is by induction on the dimension n of C. The assertion is obvious for $n = 0$, because $C = \{0\}$, $\varphi(0) = 0$ and $\alpha = 1$.

According to (4) any admissible $(n-1)$-face $s \subset \partial C$ lies in $C_n^-(k)$ and by our induction hypothesis the number $\alpha(C_n^-(k))$ of such faces is odd. Let $\alpha(S)$ denote the number of admissible $(n-1)$ – faces of an n-simplex $S \subset C$.

If S is an admissible n-simplex, clearly $\alpha(S) = 1$; while if S is not an admissible n-simplex, we have $\alpha(S) = 2$ or $\alpha(S) = 0$ according as $\varphi(S) = \{0, \dots, n-1\}$ or $\{0, \dots, n-1\} \setminus \varphi(S) \neq \emptyset$.

Hence

$$\alpha(C(k)) = \sum \alpha(S), \bmod 2. \tag{6}$$

On the other hand, an admissible $(n-1)$-face is counted exactly once or twice in $\Sigma\alpha(S)$ according as it is in the boundary of C or not. Accordingly

$$\sum \alpha(S) = \alpha(C_n^-(k)), \bmod 2, \tag{7}$$

hence

$$\alpha(C_n^-(k)) = \alpha(C(k)), \bmod 2. \tag{8}$$

But $\alpha(C_n^-(k))$ is odd. Thus $\alpha(C(k))$ is odd, too. ■

Remark. The assumption that for each $i \le n$, $f_i(I_i^-) \subset (-\infty, 0]$ and $f_i(I_i^+) \subset [0, +\infty)$, can be replaced by the Bolzano condition

$$f_i(a) \cdot f_i(b) \le 0, \quad \text{for each } i \le n \text{ and } a \in I_i^-, b \in I_i^+.$$

To verify this, let us assume that $f_i(I_i^\varepsilon) \neq \{0\}$ for each facet I_i^ε. Bolzano's condition implies that $\{f_i(I_i^-) \subset (-\infty, 0]$ and $f_i(I_i^+) \subset [0, +\infty)\}$, or $\{f_i(I_i^-) \subset [0, +\infty)$ and $f_i(I_i^+) \subset (-\infty, 0]\}$. Let us put $F_i := -f_i$ if $\{f_i(I_i^-) \subset [0, +\infty)$ and $f_i(I_i^+) \subset (-\infty, 0]\}$ and $F_i := f_i$ if not. Applying the Poincaré-Miranda Theorem to the map $F := (F_1, \dots, F_n)$, there is a point $c \in I^n$ such that $F(c) = 0$. It is clear that $f(c) = 0$, too. If $f_i(I_i^\varepsilon) = \{0\}$, then repeating this reasoning for the $(n-1)$-dimensional cube I_i^ε we get a point $c \in I_i^\varepsilon$ for which $f(c) = 0$.

Coincidence Theorem. *If maps $g, h\colon I^n \longrightarrow I^n$ are continuous and if $h(I_i^-) \subset I_i^-$ and $h(I_i^+) \subset I_i^+$ for each $i = 1, \dots, n$, then there exists a point c such that $g(c) = h(c)$.*

Proof: Let us put $f(x) := h(x) - g(x)$. The map f satisfies the assumptions of the Poincaré-Miranda Theorem and therefore there is a point $c \in I^n$ such that $f(c) = 0$. But this means that $g(c) = h(c)$. ■

Maps g and h that satisfy the conclusion of the Coincidence Theorem are said to have the *coincidence property*. If h is the identity map, we get

Bohl-Brouwer Fixed Point Theorem. *Any continuous map $g\colon I^n \longrightarrow I^n$ has a fixed point.*

Applying the Coincidence Theorem to constant maps $g(x) = a$, $a \in I^n$, gives the

Corollary. *A continuous map $h\colon I^n \longrightarrow I^n$ is onto if $h(I_i^-) \subset I_i^-$ and $h(I_i^+) \subset I_i^+$ for each $i = 1, \dots, n$.*

Borsuk Non-Retraction Theorem. *Let $f\colon X \longrightarrow R^n$ be a continuous map from a compact set $X \subset R^n$. If $f(x) = x$ for each $x \in \partial X$, then $X \subset f(X)$.*

Proof: Let J^n be an n-dimensional cube such that $X \cup f(X) \subset J^n$. Extend the map f to a continuous map h: $J^n \longrightarrow J^n$ such that $h(x) = x$ for each $x \in J^n \setminus X$. It is clear that $h(J_i^-) \subset J_i^-$ and $h(J_i^+) \subset J_i^+$ for each i. The preceding corollary ensures that $J^n \subset h(J^n)$, and hence $X \subset f(X)$. ■

Cube-Squeezing Theorem. *Let h: $I^n \longrightarrow X$ be a continuous map onto a metric space. If $h(\partial I^n) = X$, then for some $i = 1, \ldots, n$ the images of the i-th opposite faces have non-empty intersection, that is, $h(I_i^-) \cap h(I_i^+) \neq \emptyset$.*

Proof: Set $A_i := h(I_i^-)$ and $B_i := h(I_i^+)$ for each $i = 1, \ldots, n$. Suppose that $A_i \cap B_i = \emptyset$ for each i. Since X is a normal space, there exists a continuous map g: $X \longrightarrow R^n$, $g = (g_1, \ldots, g_n)$ such that $g_i(A_i) = \{0\}$ and $g_i(B_i) = \{1\}$ for each i. It is clear that $g(A_i) \subset I_i^-$ and $g(B_i) \subset I_i^+$, which implies that $g(X) \subset \partial I^n$. Let us put $f := g \circ h$. The map f: $I^n \longrightarrow \partial I^n$ satisfies the assumptions of the preceding corollary and therefore $f(I^n) = I^n$, which is a contradiction. ■

When $n = 3$, the theorem says that it is not possible to make a drawing of the cube I^3 in the plane so that disjoint faces of I^3 are disjoint in the drawing; see Figure 2. This remark leads to the conclusion that the Cube-Squeezing Theorem has something in common with dimension theory.

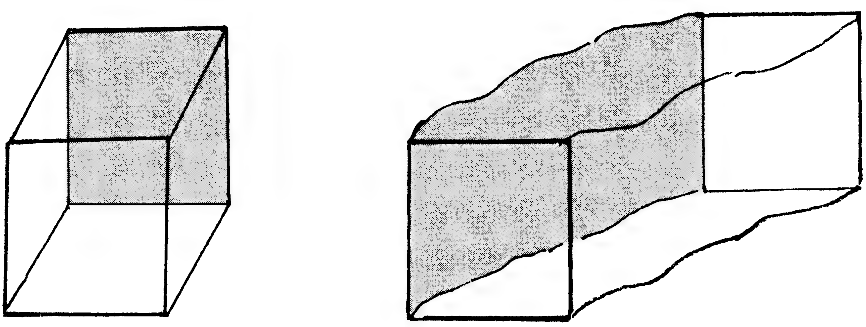


Figure 2

We shall show that the Cube-Squeezing Theorem holds whenever the set $h(I^n) \setminus h(\partial I^n)$ is "small".

Non-Squeezing Theorem. *Let h: $I^n \longrightarrow Y$ be a continuous map onto a metric space Y such that $h(I_i^-) \cap h(I_i^+) = \emptyset$ for each $i = 1, \ldots, n$. Then there exists a closed subset $A \subset Y \setminus h(\partial I^n)$ such that $\dim A \geq n$.*

Proof: Let g: $h(\partial I^n) \longrightarrow \partial I^n$ be a continuous map of the kind described in the proof of the Cube-Squeezing Theorem and suppose that dim $A < n$ for each

closed subset $A \subset Y \setminus h(\partial I^n)$. Then the map g has a continuous extension g_1: $Y \longrightarrow \partial I^n$ (see [**5**, Chapter VI]). Considering the composition of the map h with the map g_1 gives a contradiction with the Corollary. ∎

The Non-Squeezing Theorem may be thought of as a kind of Brouwer domain invariance theorem, because in the case when $Y = R^n$, dim $A = n$ if and only if the set A has non-empty interior in the space R^n.

REFERENCES

1. P. Bohl, Über die Bewegung eines mechanischen Systems in der Nähe einer Gleichgewichtung, *J. Reine Angew. Math.* 127 (1904), 179–276.
2. L. E. J. Brouwer, Beweis der Invarianz der Dimensionzahl, *Math. Ann.* 70 (1911), 161–165.
3. F. E. Browder, Fixed point theory and nonlinear problems, *Bull. Amer. Math. Soc.* 9 (1) (1983), 1–39.
4. J. H. Hadamard, Sur quelques applications de l'indice de Kronecker, Introduction to: J. Tannery, *La Théorie des Fonctions d'une Variable*, Hermann, Paris, 1910.
5. W. Hurewicz and H. Wallman, *Dimension theory*, Princeton University Press, Princeton, 1948.
6. V. Jarnik, *Bolzano and the foundations of mathematical analysis*, Society of Czechoslovak Mathematicians and Physicists, Prague, 1981.
7. B. Knaster, K. Kuratowski, and S. Mazurkiewicz, Ein Beweis des Fixpunktsatzes für n-dimensionale Simplex, *Fund. Math.* 14 (1929), 132–137.
8. C. Miranda, Un' osservazione su una teorema di Brouwer, *Boll. Unione Mat. Ital.* 3 (1940), 527.
9. H. Poincaré, Sur certaines solutions particulieres du probléme des trois corps, *C. R. Acad. Sci. Paris* 97 (1883), 251–252.
10. H. Poincaré, Sur certaines solutions particulieres du probléme des trois corps, *Bull. Astronomique* 1 (1884), 63–74.
11. H. Poincairé, Sur les courbes définies par une équation différentielle IV, *J. Math. Pures Appl.* 85 (1886), 151–217.
12. E. Sperner, Neuer Beweis für die Invarianz der Dimensionszahl und des Gebietes, *Abhandlungen aus dem Mathematisches Seminar, Hamburg Universität* 6 (1928), 265–272.

Department of Mathematics
University of Silesia
ul. Bankowa 14
40 007 Katowice, Poland
kulpa@ux2.math.us.edu.pl

Although Fermat's Last Theorem beguiles,
It appears to be suited to styles
Of diligent toilers
Like Gausses and Eulers,
So it yielded to Andrew J.'s wiles.

Contributed by Cybergeezer

Lecturing at the "Bored"

Melanie Wahlberg

Even a superficial review of the literature on collegiate mathematics education reform reveals a common thread: the imperative nature of the college mathematics student's active participation in the learning of mathematics. Active participation in the classroom may take the form of working cooperatively with classmates in small groups, spending time in the computer lab using class-related software, presenting problems or concepts on the board to classmates, or other activities germane to the subject *besides* passively listening to a lecture presented by the instructor. There is a wide range of lectures styles, but the style to which I refer is one characterized by little expectation of student engagement besides perhaps following the logic of the lesson. Because active learning in the mathematics classroom is advocated by several professional groups ([**1**], [**4**], [**5**], [**6**], [**8**]), it is certainly reasonable for mathematics instructors to reflect on their rationale for maintaining a lecture style of teaching, even part of the time. During the 1995–1996 academic year, that is exactly what I did.

I was teaching a section of reformed calculus, with a class size of forty. I had planned on extensive use of cooperative groupwork, and had accordingly divided the class into ten groups of four people each. I attempted to make the groups heterogeneous (in gender and ability) by collecting quantitative measures of the students' previous mathematical performance. Several times during the first two weeks of class, I set aside time for group work. However, due to the physical layout of the classroom (five rows of auditorium seating with twelve seats per row and an aisle up the middle) my students were having a difficult time actually communicating in their groups. Classtime initially devoted to guided discovery in the groups degenerated into work time for forty individuals, certainly not the vision I had foreseen when planning earlier that summer. To restructure this time, I tried to view my entire class as a large cooperative group, with me as the group leader. Instead of breaking the class into four-person groups as before, I would present examples of concepts we were currently studying, and pose open-ended questions to the class. For example, while developing a geometrical approach to Newton's method for finding the roots of an equation, I asked my students which solution they thought the graphical process would yield when the initial guess for the x-value was directly between two of the roots. Was there a pattern that allowed us to predict the general result? Another time when discussing inverse functions, we did an impromptu in-class exploration of which *linear* functions, and later, which *general* functions, are their own inverses. Such discussions involved much conjecturing, arguing, and scratchpaper seatwork, individually and in pairs, among the students, and were opportunities for algebraic, geometric, and numerical exploration.

Because of my class's general enthusiasm and willingness to participate, If felt we had developed a classroom atmosphere free of criticism, and open to conjecturing. The students often presented problems for each other, furthering their own mathematical communication. Although I was somewhat disappointed not to be able to watch my students develop cooperative skills, I felt that we had made the

best of the situation, and reached a workable compromise: modified traditional lecturing in a nontraditional atmosphere.

This arrangement continued in my class for about eight weeks until we came to the point where the students were ready to see the Fundamental Theorem of Calculus, the culmination of much study of differentiation and some study of integration, principally from a conceptual viewpoint. Primarily because I was so enchanted with the textbook's intuitive approach (we were using [**2**], produced by the Consortium based at Harvard), I decided to use the last fifteen minutes of class to present the proof of the theorem, instead of the usual timeshare between examples and discussion. After all, we had certainly developed the necessary machinery, and I thought exposure to the underpinnings of this grand relationship would benefit all of us. My thoughtful reflection and development of the lecture would certainly make the connections clear to everyone.

Proving the Fundamental Theorem of Calculus to my class was one of the highlights of teaching that semester. This was to be the only theorem whose proof I presented, but I felt its phenomenal usefulness merited a little extra attention. I derived a great deal of satisfaction from my presentation of the proof, and was quite happy with the support of the textbook as well. The class was relatively quiet during that period, but I felt the students just *must* have understood and appreciated the lecture. The examples and problems the students had worked (or at least seen) were an almost perfect framework in which to nestle this new theorem, and I myself looked forward to working the homework problems that required its application.

The next day I was surprised by the unusually large number of students who appeared during my office hour. My small office began to overflow, so we moved to a classroom with a blackboard to better accommodate ourselves. It appeared that the students were having a hard time even starting the homework problems I had assigned for practice, let alone knowing when to apply their grand new theorem. Together, my students and I worked through three or four of the problems, appealing frequently to the theorem. At one point, one of my students cried out, "I'm not sure I even believe that relationship... how would you *prove* something like that?" The other students then began talking among themselves: "The proof is that thing she did yesterday at the end of class", and "That was so confusing!" and worst of all, "Were you *listening* to what she was talking about? I was staring out the window the whole time!" I felt a strong sense of disappointment. After all, I had enjoyed giving the presentation so much! At this point, I seriously began to consider my reasons for standing up at the board.

The last straw came a couple of weeks later, when I gave my class a midsemester course evaluation, to be completed anonymously. My students were handing in high quality work, and class participation was greater than in any class I had taught before, but I was interested in their impressions. One student wrote, "I wish we had more time for discussion among ourselves, instead of watching the instructor up at the board." (sic) This very telling slip on my student's part, combined with the earlier feedback, made me ask myself if I could ever stand up at the "bored" without my students mentally shutting down.

Looking for an answer to this question forced me to ask myself another: "What is it that keeps me at the board, *telling* my students about mathematics, with seamless lectures and well-chosen examples?" For an additional perspective, I posed this question to a colleague of mine. He and I quickly generated a chalkboard full of reasons why we continue to lecture, even after frequent and positive experiences with alternative forms of teaching. Upon further reflection, I

discovered that the reasons we listed fell into three categories: reasons lecturing seems to benefit the instructor directly, reasons the instructor thinks lecturing benefits the students, and reasons students prefer the instructor to lecture; these three areas were not disjoint. I organized and sorted the list, and went on to respond mentally to each reason we had listed. Following is an organized list of these reasons for lecturing and "reflections" responding to each reason. To preserve the spirit in which these ideas developed two voices (a point and a counterpoint) will be heard throughout this discussion. The first voice reflects my initial tendencies, and, based on my own mathematics classroom training and experience, advocates the lecture method. The second voice counters the first, suggesting ways of maintaining mathematical standards while attempting to involve students more actively in the learning process.

REASONS THAT APPEAR TO BENEFIT THE INSTRUCTOR DIRECTLY. First, lecturing gives me a feeling of control. If I stand up at the board day after day and "present" or "cover" the material, I know exactly what my students have seen. I am quite able to keep up with any department-mandated syllabus, and can expose my students to exactly the techniques, examples, and ideas I want them to think about. The pace might not be quite right for each student, but I am confident that the middle third is reasonably challenged, and I am more than willing to interrupt my monologue to answer questions.

A second reason I continue to lecture is that is assuages any guilt I might feel after using alternative methods for a period of time. What if they are not really "getting it" in their cooperative groups, or their interactive sessions in the computer lab? If I *tell* them the rules, and *show* them the examples, the responsibility for their learning somehow transfers to their shoulders. Now that I've said it, it is fair game on homework and exams. If my students aren't able to demonstrate mastery of the material when I assess them, the burden lies on them. They must not have been listening when I presented this concept?

My third reason for lecturing is more personal than the first two. I get tremendous satisfaction from preparing and presenting a well-laid out elegant set of accessible ideas, each of which builds on the one before. After all, this is material I love, and a favorite way to share it is to lay it out neatly for students. In addition, each time I prepare a topic to be presented for my students' eyes I gain additional mathematical insight. It is very easy for me to generate my own enthusiasm and come back to class day after day when I am "presenting."

My last reasons really falls into all three categories: I am comfortable at the board, and my students are comfortable (perhaps too much so) when I am at the board as well. It is more challenging for all of us when I have them work in groups or present problems for each other.

REFLECTIONS: Let's first consider the idea of control. While it is true that by lecturing, I control the material being presented, I certainly have no control over what the students choose to listen to, and even less over what they choose to study. If instead I employ alternative methods of teaching, such as cooperative group-work, I really am not relinquishing any *valuable* control, and instead gain the possibility of more actively engaging the students in the learning process. On one hand, it is certainly possible for a student to be an active participant in the mathematical process of listening to a lecture. He or she may pose questions, anticipate examples or counterexamples of concepts being presented, or make educated guesses about the next step of the proof. However, my experience with

randomly calling on students during a lecture would imply that such engagement is not the norm. Cooperative groupwork offers instructors an alternative classroom modality with a chance of greater student participation; it is much more difficult to "hide" in a group of four than in a classroom of forty.

I might let go of my second concern, namely the guilt associated with letting students pose and verify their own conjectures, when I realize that the lecture method ensures the coverage of, but not necessarily the learning of material at hand. As long as the group activities (lab work, projects, etc.) are designed to include the students' reflecting on what it is they have discovered, coupled with timely feedback from me, I may feel just as confident the material has been duly "covered", and perhaps with greater student understanding, due to their participation.

An example of an activity specifically designed to determine whether or not students were making their own connections outside of class was an expository paper requiring students to reconcile the formal definition of the derivative of a function at a point with their personal conceptual understanding. I provided only a few guidelines: students were to use graphical, numerical, and algebraic approaches in their explanations and were to enumerate at least two distinct applications of the derivative of a function at a point. Their audience was a (fictitious) fellow student who had missed the pertinent days of class.

Completing this writing assignment provided a basis of understanding that many of my students referred to repeatedly through both semesters of the calculus sequence. The struggle to gather, refine, and articulate their thoughts encouraged them to expand what Vinner [9] terms the student's *concept image* (that is, the nonverbal entity, correct or note, that the student associates with the concept name) until it became compatible with the *concept definition* of a derivative at a point. Thus, I did not spend class time trying to deliver a lecture tailored to fill in the gaps of my students' individual and collective understanding of derivative. Instead, I developed a task that both enabled students to make their own connections and provided a window into the understanding each had constructed of algebraic, geometric and numerical relationships.

In addressing the third reason for lecturing, I must first admit that the satisfaction I get from preparing a lecture is undeniable, and stems from at least two components: assembling my mathematical ideas into a coherent structure, and the resulting deeper understanding. However, while preparing for a class session that more actively involves a greater portion of the students, I may experience both of these components by changing my expectations for the class period, and devising supportive materials accordingly. For example, Monk and Finkel [3] suggest putting the creative energy usually reserved for planning lectures into constructing a sequence of probing, converging questions that elicit critical student thinking. These questions might then be answered by the students after they have studied a mathematical concept; their answers can provide material for group discussion.

Finally, while the comfort of the well-known lecture method is indisputable, it is a comfort that may breed complacency among the students and in the instructor. Thus it is not necessarily a benefit. Employing alternative methods of teaching certainly forces instructors and students outside their "comfort zones." However, Reynolds et al. [7] say as long as the instructor is "... an enthusiastic advocate of working cooperatively...," students tend to adapt to cooperative groupwork to the point that the class structure does not interfere unduly with their learning.

REASONS FOR LECTURING THAT APPEAR TO BENEFIT THE STUDENTS. These reasons are strictly from an instructor's point of view, the student's perspec-

tive is discussed later. One driving force behind the lecture method has been that it provides quick exposure to a body of knowledge to many people at one time. With it, I might summarize years of investigative work in a fifty-minute session. Thus I can show my students a larger amount of material in one semester.

The second benefit for students goes hand-in-hand with the preceding one: a well-laid out lecture gives students, particularly upperclassmen, exposure to the potential beauty and elegance of mathematics, and provides a framework in which to place the problems and proofs they are working. When students have access to such a lecture, they might get as much pleasure from listening as I do from telling.

Finally, in this category is the fact that lecturing demonstrates to students at all levels one way in which to convey mathematical ideas, both verbally and symbolically. Day after day, students can listen and watch a person speak and write mathematics. There is an unstated assumption that students are not only learning the mathematical content being presented, but are somehow internalizing the style in which a mathematician communicates mathematics.

REFLECTIONS: I will rebut all three reasons of this category simultaneously. The weakness of each argument (from the standpoint of my calculus class) is the same: the goals are very lofty, but they are not being realized. Lecturing would be a remarkable pedagogical tool if students were able to *internalize* "years of investigative work," *appreciate* the "potential beauty and elegance of mathematics," and *learn* to "convey mathematical ideas, both verbally and symbolically," after listening to the teacher standing at the board. However, my calculus class demonstrated through their inability to connect with the Fundamental Theorem of Calculus (and in other instances, such as applying the chain rule, or carrying out numerical algorithms) that this often is not the case.

REASONS THAT STUDENTS PREFER AN INSTRUCTOR TO LECTURE. As much as students complain of the tedium associated with listening to an instructor for a significant portion of the class period, there still are reasons students may prefer this format to another. The first is a natural consequence of the climate established in the typical lecture-format classroom: when the teacher presents exactly the material that is needed for successful completion of the course, there is no need for the students to read the textbook. For the students, the textbook is then an example (template) reference for homework problems, and a source for teacher-assigned exercises. A corollary to this is that when the teacher exclusively employs the lecture method, it sends the very clear message to the students that whatever is important will be covered in class, so there is no need to sort out for oneself the major concepts from the supportive material. In particular, an instructor in this setting would be breaking an unwritten rule to test students on an idea not explicitly covered in lecture.

Finally, students like an instructor to lecture for the simple reason that it is much more restful, both physically and mentally, than the more active group effort required to solve a problem or individual effort required to present a problem on the board. Depending on the classroom dynamic, students could go through the whole semester without speaking in their math class if they never ventured to ask a question.

REFLECTIONS: The reasons from the students' perspective are probably the easiest category to counter. If students receive the message that the prose of the text is superfluous (don't read it), and further that the teacher will implicitly make

plain what should be studied (what is on the test), they cannot be expected to develop mathematical autonomy. Similarly, for many students, the role of listening to a lecture, no matter how conceptually challenging, does not imply the development of ownership of the mathematical ideas being presented.

SUMMARY. Based on the collective experiences of this calculus class and on other courses I have taken and taught, I claim that the classroom with lecture as the only mode for instruction is obsolete. Lecture tends to minimize the opportunity for engagement of the learner with mathematics. Fortunately, making a transition from a classroom atmosphere having exclusively the traditional lecture format to one characterized by a blend of lecturing and alternative learning methods need not imply a compromise. I have argued that having students present problems to the whole class, for instance, does not lower academic standards but rather demands and develops a higher level of mathematical communication Instructors employing cooperative groupwork do not relinquish control over the curriculum; they instead recreate the atmosphere in which students learn the mathematics of a curriculum determined by the instructor. Thus, instructors may simultaneously maintain mathematical standards and engage a greater fraction of their students with the material, an activity not explicitly demanded in a classroom where the instructor is the sole resource for knowledge. Lecturing should be limited to situations where students are at least partially engaged due to the climate the instructor has developed, and should be used to introduce students to some basic concepts as a springboard for their students' own discoveries. This extends the power of the lecture, allowing the instructor to maximize its potential and to minimize lecturing at the bored.

REFERENCES

1. Fisher, Naomi D. (ed.) *Changing the Culture: Mathematics Education in the Research Community. Issues in Mathematics Education* **5**, American Mathematical Society 1995.
2. Hughes-Hallett, Deborah and Andrew Gleason, et al. *Calculus*. New York, New York. John Wiley & Sons, Inc. 1994.
3. Monk, Steven G. and Donald Finkel. *Teachers and Learning Groups: Dissolution of the Atlas Complex*. In *Learning in Groups*. Jossey-Bass 1983, pp. 82–97.
4. National Council of Teachers of Mathematics. *Curriculum and Evaluation Standards for School Mathematics*. NCTM Reston, VA. 1989.
5. National Research Council. *Everybody Counts: A Report to the Nation*. Washington D.C.: National Academy Press 1989.
6. National Research Council. *Moving Beyond Myths: Research in Undergraduate Mathematics*. Washington, D.C.: National Academy Press 1991.
7. Reynolds, Barbara E., et al. (eds.) *A Practical Guide to Cooperative Learning in the Undergraduate Mathematics Classroom*. MAA Notes **37**, Mathematical Association of America 1995.
8. Roberts, A. Wayne (ed.) *Calculus: The Dynamics of Change*. MAA Notes **39**, Mathematical Association of America 1996.
9. Vinner, Shlomo. *The Role of Definitions in Teaching and Learning*. In *Advanced Mathematical Thinking* (ed. David Tall). Kluwer Academic Publishers, 1991.

Department of Mathematics & Statistics
Western Michigan University
Kalamazoo, MI 49008
mking@kzoo.edu

NOTES

Edited by **Jimmie D. Lawson**

A Generalization of Wolstenholme's Theorem

M. Bayat

In this article we prove a generalization of Wolstenholme's theorem [2] with a simpler proof using group theory and number theory.

Theorem 1. (Wolstenholme). *If $p \geq 3$ is prime, then the numerator of the fraction $1 + \frac{1}{2} + \cdots + \frac{1}{p-1}$ is divisible by p^2.*

Proof: First observe that

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{p-1}$$

$$= \left(1 + \frac{1}{p-1}\right) + \left(\frac{1}{2} + \frac{1}{p-2}\right) + \cdots + \left(\frac{1}{\frac{p-1}{2}} + \frac{1}{\frac{p+1}{2}}\right)$$

$$= p\left(\frac{1}{1(p-1)} + \frac{1}{2(p-2)} + \cdots + \frac{1}{\left(\frac{p-1}{2}\right)\left(\frac{p+1}{2}\right)}\right)$$

$$= p\frac{A}{(p-1)!},$$

where

$$A = \frac{(p-1)!}{1(p-1)} + \frac{(p-1)!}{2(p-2)} + \cdots + \frac{(p-1)!}{\left(\frac{p-1}{2}\right)\left(\frac{p+1}{2}\right)}.$$

In the following it is shown that A is divisible by p. Consider now the multiplicative Z_p-group consisting of all non-zero congruence classes mod p. We choose $\bar{r}$ in the Z_p-group corresponding to $(p-1)!/i(p-i)$, i.e.,

$$r \equiv \frac{(p-1)!}{i(p-i)} \pmod{p} \quad \left(1 \leq \frac{p-1}{2}\right).$$

Thus

$$ri(p-i) \equiv (p-1)! \pmod{p} \quad \left(1 \le i \le \frac{p-1}{2}\right).$$

According to Wilson's theorem [**1**], $(p-1)! \equiv -1 \pmod{p}$, and so

$$-ri^2 \equiv -1 \pmod{p} \quad \left(1 \le i \le \frac{p-1}{2}\right).$$

In Z_p this means that $\overline{i^2}$ is the inverse of $\bar{r}$:

$$\bar{r} = \left(\overline{i^2}\right)^{-1} \quad \text{or} \quad \bar{r} = \left(\overline{i^{-1}}\right)^2 \qquad \left(1 \le i \le \frac{p-1}{2}\right).$$

Since

$$\overline{i^2} = \overline{(p-i)}^2 \qquad \left(1 \le i \le \frac{p-1}{2}\right),$$

every square of Z_p arises as a square from only half the elements of Z_p, $\{\bar{1}, \bar{2}, \dots, \overline{\frac{p-1}{2}}\}$. Furthermore, the squaring map is injective on this set. Since the inverse map is bijective on the square elements, every $\bar{r}$ is equal to the square of one of the numbers, $\bar{1}, \bar{2}, \dots, \overline{\frac{p-1}{2}}$; hence

$$A = 1^2 + 2^2 + \cdots + \left(\frac{p-1}{2}\right)^2 = \left(\frac{p-1}{2}\right)\left(\frac{p+1}{2}\right)\left(\frac{p}{6}\right) \equiv 0 \pmod{p}$$

and this completes the proof of theorem.

Theorem 2. *If $p \ge 3$, then the numerator of* $1 + \frac{1}{2^2} + \cdots + \frac{1}{(p-1)^2}$ *is divisible by p.*

Proof: The numerator of the fraction is

$$A = \frac{(p-1)!^2}{1} + \frac{(p-1)!^2}{2^2} + \cdots + \frac{(p-1)!^2}{(p-1)^2}.$$

First of all we calculate $(p-1)!^2/i^2$ for $1 \le i \le p-1$. Let i be arbitrary and $1 \le i \le p-1$. There exists $\bar{r}$ in Z_p such that

$$r \equiv \frac{(p-1)!^2}{i^2} \pmod{p} \quad (1 \le i \le p-1),$$

and thus

$$ri^2 \equiv 1 \pmod{p}.$$

In other words, $\overline{i^2}$ is the inverse of $\bar{r}$ in Z_p. Indeed

$$\bar{r} = \left(\bar{i}^{-1}\right)^2 = \left(\overline{p-i}^{-1}\right)^2 \quad (1 \le i \le p-1).$$

Therefore, $\bar{r}$ is one of the following:

$$\bar{1}^2, \bar{2}^2, \ldots, \left(\overline{\frac{p-1}{2}}\right)^2.$$

Hence

$$A \equiv 2\left(1^2 + 2^2 + \cdots + \left(\frac{p-1}{2}\right)^2\right) \equiv 0 \pmod{p},$$

which gives $p|a$.

Theorem 3. *Let p be prime and $k \in N$ such that $2k \leq p-1$. Then*
(i) the numerator of the fraction

$$1 + \frac{1}{2^{2k-1}} + \cdots + \frac{1}{(p-1)^{2k-1}}$$

is divisible by p^2.
(ii) The numerator of the fraction

$$1 + \frac{1}{2^{2k}} + \cdots + \frac{1}{(p-1)^{2k}}$$

is divisible by p.

Proof: (i) The numerator of the fraction is

$$A = \frac{(p-1)!^{2k-1}}{1^{2k-1}} + \frac{(p-1)!^{2k-1}}{2^{2k-1}} + \cdots + \frac{(p-1)!^{2k-1}}{(p-1)^{2k-1}}.$$

To each element of A we can associate an element of $\bar{r} \in Z_p$. In fact,

$$r \equiv \frac{(p-1)!^{2k-1}}{i^{2k-1}} \pmod{p} \quad (1 \leq i \leq p-1)$$

satisfies $ri^{2k-1} \equiv -1 \pmod{p}$. This means that in Z_p, $\bar{r} = (\overline{-i}^{2k-1})^{-1}$ $(1 \leq i \leq p-1)$, and, since the inverse is unique, $-\bar{r}$ is one of the 2^{2k-1}-powers of $\bar{1}, \bar{2}, \ldots, \overline{(p-1)}$. Hence, the numerator of the fraction is

$$A \equiv -\left(1 + 2^{2k-1} + \cdots + (p-1)^{2k-1}\right) \pmod{p} \tag{1}$$

In order to calculate the right-hand side of (1), we use Euler's formula **[1]**

$$1^m + 2^m + \cdots + (n-1)^m = \sum_{r=0}^{m} \frac{1}{m+1-r}\binom{m}{r} n^{m+1-r}\beta_r,$$

where the β_r are Bernoulli's numbers. Indeed $\beta_0 = 1$, $\beta_1 = -\frac{1}{2}$, $\beta_{2k+1} = 0$, and $\beta_{2k} = (-1)^{k-1}B_k$ for $k \geq 1$, where the B_k are the coefficients of x^{2k} in the series expansion

$$\frac{x}{\exp - 1} = \left(1 - \frac{1}{2}x + \frac{B_1}{2!}x^4 + \frac{B_3}{6!}x^6 - \cdots\right).$$

This yields

$$1^{2k-1} + 2^{2k-1} + \cdots + (p-1)^{2k-1} = \sum_{r=0}^{2k-1} \frac{1}{2k-r}\binom{2k-1}{r} p^{2k-r}\beta_r,$$

and so

$$2k\left(1^{2k-1}+2^{2k-1}+\cdots+(p-1)^{2k-1}\right)=\sum_{r=0}^{2k-1}\frac{2k}{2k-r}\binom{2k-1}{r}p^{2k-r}\beta_r.$$

The right-hand side is divisible by p^2 since $\beta_{2k-1}=0$, and hence

$$p^2|2k\left(1^{2k-1}+2^{2k-1}+\cdots+(p-1)^{2k-1}\right).$$

Since $2k\le p-1$, this implies $p^2|A$, and this completes the proof of (i).The proof of (ii) is similar to (i).

Theorem 4. *Let p be a prime and let $k\in N$ such that $2k\le p-1$. Then*

(i) *The numerator of the fraction $\Sigma_{i=1}^{p^n-1}1/i^{2k-1}$, $[(i,p)=1]$ is divisible by p^{n+1}.*
(ii) *The numerator of the fraction $\Sigma_{i=1}^{p^n-1}1/i^{2k}$, $[(i,p)=1]$ is divisible by p^n.*

Proof: (i) According to Theorem 3 the numerator of each of the following numbers is divisible by p^2:

$$A_1=\frac{1}{1^{2k-1}}+\frac{1}{2^{2k-1}}+\cdots+\frac{1}{(p-1)^{2k-1}},$$

$$A_2=\frac{1}{(p-1)^{2k-1}}+\frac{1}{(p-2)^{2k-1}}+\cdots+\frac{1}{(2p-1)^{2k-1}},$$

$$\vdots$$

$$A_{p^{n-1}}=\frac{1}{(p^n-p+1)^{2k-1}}+\frac{1}{(p^n-p+2)^{2k-1}}+\cdots+\frac{1}{(p^n-1)^{2k-1}}.$$

A proof similar to the one given in Theorem 3 shows that the numerator of each of these fractions is equivalent to $1^{2k-1}+2^{2k-1}+\cdots+(p-1)^{2k-1}$ (mod p). Since there are p^{n-1} of these fractions we have

$$p^{(n-1)}\left(1^{2k-1}+2^{2k-1}+\cdots+(p-1)^{2k-1}\right)\equiv 0\pmod{p^{(n+1)}}.$$

and the numerator of $\Sigma_{r=0}^{p^{n-1}}A_i$ is divisible by p^{n+1}. This completes the proof of (i). A similar argument proves (ii).

ACKNOWLEDGMENT. The author wishes to thank Professor Y. Sobouti for a helpful mix of prodding and advice.

REFERENCES

1. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Fourth Edition, Clarendon Press, Oxford, 1960.
2. Emre Alken, Variations on Wolstenholme's theorem, *Amer. Math. Monthly* 101 (1994), 1001–1004.

Institute for Advanced Studies in Basic Sciences
P.O. Box 45195-159
Gaveh Zang
Zanjan, IRAN
Iasbsgz1@Rose.ipm.ac.ir

A Note on the Mean Value Theorem for Integrals

Zhang Bao-lin

The purpose of this note is to extend a result by Bernard Jacobson [**1**] concerning the mean value theorem for integrals.

Theorem 1. *If the function f is continuous on the interval $[a, b]$, then there is a number c such that $a < c < b$ and*

$$\int_a^b f(x)\,dx = f(c)(b-a). \tag{1}$$

If $x \in (a, b)$ then by Theorem 1 applied to the interval $[a, x]$, it is possible to choose a number c_x $(a < c_x < x)$ as a function of x on (a, b) such that

$$\int_a^x f(t)\,dt = f(c_x)(x-a). \tag{2}$$

Jacobson studied the behavior of c_x as x approaches a, and proved the following result:

Theorem 2. *Suppose the function f is continuous on the interval $[a, b]$ and is differentiable at a with $f'(a) \neq 0$. If c_x is given in the mean value formula* (2) *for integrals, then*

$$\lim_{x \to a} \frac{c_x - a}{x - a} = \frac{1}{2}. \tag{3}$$

We consider the case $f'(a) = 0$ in Theorem 2.

Theorem 3. *Suppose the function f is continuous on the interval $[a, b]$ and is twice-differentiable at a with $f'(a) = 0$, $f''(a) \neq 0$. If c_x is given in the mean value formula* (2) *for integrals, then*

$$\lim_{x \to a} \frac{c_x - a}{x - a} = \frac{1}{\sqrt{3}}. \tag{4}$$

Proof: Consider Taylor's expansion

$$f(t) = f(a) + f''(a)\frac{(t-a)^2}{2} + \epsilon(t)(t-a)^2 \tag{5}$$

where $\epsilon(t) \to 0$ as $t \to a$. Integrate (5) from a to x, and obtain

$$\int_a^x f(t)\,dt = f(a)(x-a) + f''(a)\frac{(x-a)^3}{3!} + \int_a^x \epsilon(t)(t-a)^2\,dt. \tag{6}$$

On the other hand, equation (5) evaluated at c_x $(a < c_x < b)$ yields

$$f(c_x) = f(a) + f''(a)\frac{(c_x - a)^2}{2} + \gamma(c_x)(c_x - a)^2, \tag{7}$$

where $\gamma(c_x) \to 0$ as $c_x \to a$. Thus

$$f(c_x)(x - a) = f(a)(x - a) + f''(a)\frac{(c_x - a)^2}{2}(x - a) + \gamma(c_x)(c_x - a)^2(x - a). \tag{8}$$

From (2), (6), and (8) we have

$$f''(a)(x - a)^3 + 6\int_a^x \epsilon(t)(t - a)^2\,dt = f''(a)3(c_x - a)^2(x - a) + 6\gamma(c_x)(c_x - a)^2(x - a). \tag{9}$$

It is easy to see that

$$\lim_{x \to a} \frac{\int_a^x \epsilon(t)(t - a)^2\,dt}{(x - a)^3} = 0, \quad \lim_{x \to a} \frac{\gamma(c_x)(c_x - a)^2}{(x - a)^2} = 0.$$

From the condition $f''(a) \neq 0$ and from (9) we obtain

$$\lim_{x \to a} \left(\frac{c_x - a}{x - a}\right)^2 = \frac{1}{3}.$$

In a similar way one can establish the following more general result.

Theorem 4. *Suppose the function f is continuous on the interval $[a, b]$, and is k times differentiable at a with $f^{(i)}(a) = 0$ $(i = 1, 2, \ldots, k - 1)$, $f^{(k)}(a) \neq 0$. If c_x is given in the mean value formula* (2) *for integrals, then*

$$\lim_{x \to a} \frac{c_x - a}{x - a} = \frac{1}{\sqrt[k]{k + 1}}.$$

REFERENCE

1. Bernard Jacobson, On the Mean Value Theorem for Integrals, *Amer. Math. Monthly* 89 (1982), 300–301.

Laboratory of Computational Physics
IAPCM, P.O. Box 8009, Beijing
China
jslcp@iapcm.ac.cn

UNSOLVED PROBLEMS

Edited by **Richard Nowakowski**

In this department the MONTHLY presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Nowakowski, Department of Mathematics & Statistics & Computing Science, Dalhousie University, Halifax NS, Canada B3H 3J5, rjn@cs.dal.ca

When Is There a Latin Power Set?

J. Dénes

In [**6**] we find that the images of 0123 under the permutations (0)(123), (1)(032), (2)(013), (3)(021) and those under their squares (0)(132), (1)(023), (2)(031), (3)(012) respectively form the Latin squares

$$L = \begin{matrix} 0 & 2 & 3 & 1 \\ 3 & 1 & 0 & 2 \\ 1 & 3 & 2 & 0 \\ 2 & 0 & 1 & 3 \end{matrix} \qquad L^2 = \begin{matrix} 0 & 3 & 1 & 2 \\ 2 & 1 & 3 & 0 \\ 3 & 0 & 2 & 1 \\ 1 & 2 & 0 & 3 \end{matrix}$$

In general, a set of n permutations $\alpha_0, \alpha_1, \dots, \alpha_{n-1}$ on a set S of n symbols form a Latin square in this way just if they are *sharply transitive*, i.e., if, for all $i \in S$ and $0 \le j < k \le n-1$, $\alpha_j(i) \ne \alpha_k(i)$. It may happen that the set of squares $\alpha_0^2, \alpha_1^2, \dots, \alpha_{n-1}^2$ are also sharply transitive, so that they too form a Latin square L^2, which is necessarily orthogonal to L. If each set of k-th powers $\alpha_0^k, \alpha_1^k, \dots, \alpha_{n-1}^k$ for $1 \le k \le m$ is sharply transitive, then the associated Latin squares $L, L^2, \dots, L^m$ form a *Latin power set*, which is a mutually orthogonal set; see [**4**] or [**3**, p. 231].

Conjecture 1 [4]. If $n \ge 4$ and $\ne 6$, then there exists a Latin power set containing two Latin squares of order n.

If true, this would give another refutation of Euler's 1782 conjecture that no pairs of orthogonal Latin squares of order n exist for $n \equiv 2 \bmod 4$. Euler's conjecture was, of course, disproved in [**2**]; see also [**3**, §11.2] and it is now known that for $n \ne 2$ or 6 there exists a pair of orthogonal Latin squares of order n. Conjecture 1 is known to be true for $7 \le n \le 50$ and for all larger n except possibly those of shape $6k + 2$; see [**4**] for details.

Keedwell ([**5**] and see [**3**, §7.1]) used a special set of sharply transitive permutations to obtain orthogonal Latin squares, which suggests the following

Conjecture 2. For every even n, $n \ge 8$, there exists a set of permutations each of which fixes one symbol while the other $n-1$ symbols form a single cycle, such that the associated L and L^2 are Latin squares.

Our opening example is of this type. Conjecture 2 is true when n is a prime power q. In this case take $\alpha_0 = (0)(1, \sigma, \sigma^2, \ldots, \sigma^{q-2})$ where σ is a primitive root of the Galois field GF$[q]$ and

$$\alpha_{i+1} = (\sigma^i)(1 + \sigma^i, \sigma + \sigma^i, \ldots, \sigma^{q-2} + \sigma^i)$$

for $0 \le i \le q - 2$ (see [**8**]). More generally, we can construct a Latin power set with as many as h members if there is a group G of order n that is *R_h-sequenceable.*

We say that a group $(G, +)$ is *R-sequenceable* (R_1-sequenceable) if its elements $a_0 = 0, a_1, \ldots, a_{n-1}$ can be ordered in such a way that the partial sums $b_0 = a_0$, $b_1 = a_0 + a_1$, $b_2 = a_0 + a_1 + a_2, \ldots, b_{n-2} = a_0 + a_1 + \cdots + a_{n-2}$ are all different and $b_{n-1} = a_0 + a_1 + \cdots + a_{n-1} = b_0 = 0$. This means that $b_1 - b_0$, $b_2 - b_1, \ldots, b_{n-2} - b_{n-3}$, $b_0 - b_{n-2}$ are all distinct and so coincide with the non-identity elements of G. If further, for each $j, j = 2, 3, \ldots, h$ the set of differences $b_{i+j} - b_i$, $i = 0, 1, \ldots, n - 2$ also are all distinct, where suffixes are taken modulo $n - 1$, we say that $(G, +)$ is *R_h-sequenceable.*

To construct a power set with h members we take $\alpha_0 = (c)(b_0, b_1, \ldots, b_{n-2})$ where the b_i are the partial sums for the R_h-sequencing and c is the element that does not occur among the partial sums, and

$$\alpha_j = (c + a_j)(b_0 + a_j, b_1 + a_j, \ldots, b_{n-2} + a_j)$$

for each group element a_j, $1 \le j \le n - 1$. Then the permutations $\alpha_0^k, \alpha_1^k, \ldots, \alpha_{n-1}^k$, form a sharply transitive set for $k = 1, 2, \ldots, h$ and so we can construct a Latin power set with h members. Note that the Galois field construction is a special case since the additive group of GF$[q]$ is R_{q-2}-sequenceable (see [**7**]). Note also that the Latin squares constructed by this method, i.e., the method of Conjecture 2, are all idempotent. It has been shown in [**5**], see also [**3**, §7.4], that if the requirement of idempotency is dropped then an R_h-sequenceable group of order n permits construction of at least $h + 1$ pairwise orthogonal Latin squares.

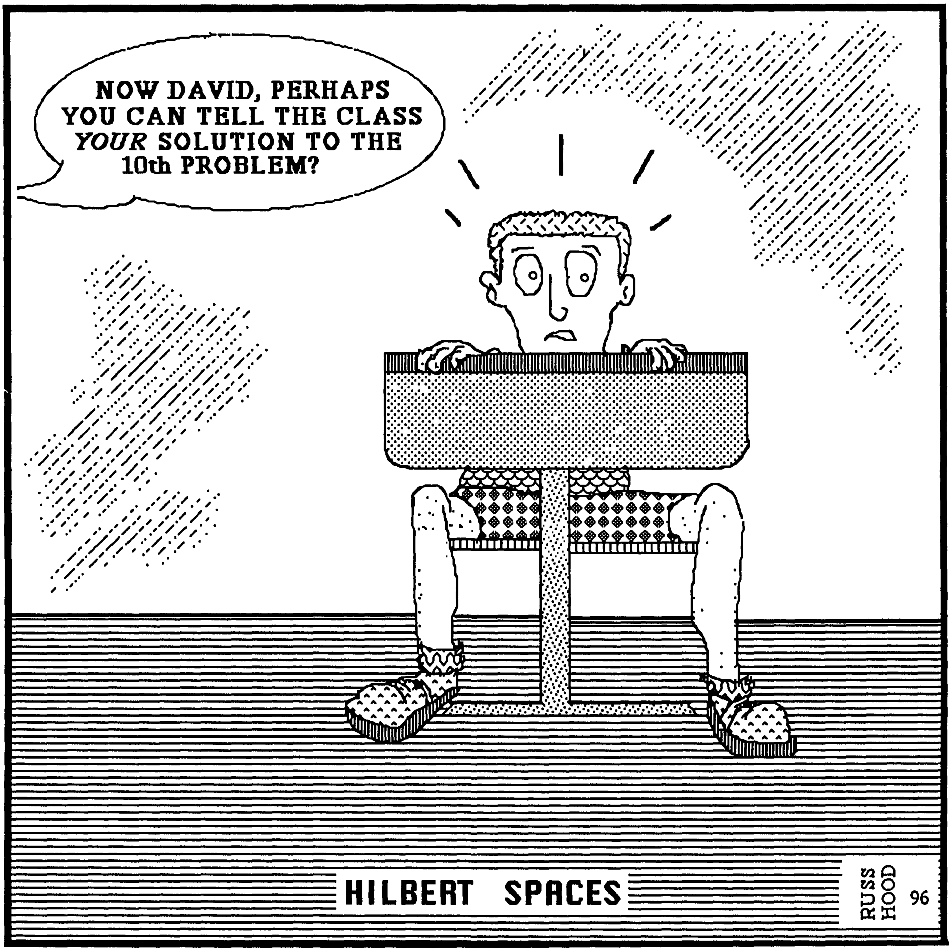
For the first open case, $n = 10$, the construction using an R_2-sequencing does not work, because neither the cyclic nor the dihedral group of any singly-even order is R-sequenceable [**7**] (nor, *a fortiori*, is it R_2-sequenceable). However, the existence of a set of permutations of the type postulated in Conjecture 2 is equivalent to the existence of a 2-fold perfect resolvable $(10, 9, 1)$-Mendelsohn design [**8**, p. 86] and the possible existence of a suitable "irregular" (non-cyclically generated) design is not ruled out. Another quite different construction of a set of h, but not of $h + 1$, pairwise orthogonal idempotent Latin squares from a set of permutations of the above kind, or from any resolvable h-fold perfect $(n, h + 1, 1)$-Mendelsohn design, is given in [**1**].

REFERENCES

1. F. E. Bennett, E. Mendelsohn and N. S. Mendelsohn, Resolvable perfect cyclic designs, *J. Combin. Theory Ser. A* 29 (1980) 142–150.
2. R. C. Bose, S. S. Shrikande, and E. T. Parker, Further results on the construction of mutually orthogonal Latin squares and the falsity of Euler's conjecture, *Canad. J. Math.* 12 (1960) 189–203.
3. J. Dénes and A. D. Keedwell, *Latin Squares and their Applications*, Academic Press, New York; Akadémiai Kiadó, Budapest; English Universities Press, London, 1974.
4. J. Dénes, G. L. Mullen, and S. J. Suchower, A note on power sets of Latin squares, *J. Combin. Math. Combin. Comput.* 16 (1994) 27–31.
5. A. D. Keedwell, On orthogonal Latin squares and a class of neofields, *Rend. Mat. e Appl.* (5) 25 (1966) 519–561.

6. A. D. Keedwell, Concerning the existence of triples of pairwise almost orthogonal 10 × 10 Latin squares, *Ars Combin.* 9 (1980) 3–10.
7. A. D. Keedwell, On R-sequenceability and R_h-sequenceability of groups, *Combinatorics* '81 (*Rome* 1981) 535–548, *North-Holland Math. Stud.* 78, North-Holland, Amsterdam–New York, 1983.
8. A. D. Keedwell, Circuit designs and Latin squares, *Ars Combin.*, 17 (1984) 79–90.

Csaba u.10
H-1122 Budapest
Hungary



Contributed by Russ Hood, Rio Linda, CA

PROBLEMS AND SOLUTIONS

Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**
with the collaboration of Paul T. Bateman, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Underwood Dudley, Michael A. Filaseta, Ira M. Gessel, Bart Goddard, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Robert Israel, Murray S. Klamkin, Daniel J. Kleitman, Fred Kochman, Frederick W. Luttmann, Frank B. Miles, M. J. Pelling, Richard Pfiefer, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before November 30, 1997; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.*

PROBLEMS

The problems in the May issue should have been numbered from **10592** through **10598** but were inadvertently numbered from **10585** through **10591**, duplicating the numbers from the April issue. We ask our readers to use the intended numbers **10592–10598** for solutions, reference, and indexing for problems in the May issue.

10599. *Proposed by Fred Galvin, University of Kansas, Lawrence, KS.* Let $x_1, \dots, x_m$ and $y_1, \dots, y_n$ be nonnegative numbers and let (a_{ij}) be an $m \times n$ matrix of nonnegative numbers with at least one nonzero entry in each row. Suppose that the inequality $\sum_{h=1}^{m} a_{hj}x_h \le \sum_{k=1}^{n} a_{ik}y_k$ holds whenever $a_{ij} > 0$. Show that $\sum_{i=1}^{m} x_i \le \sum_{j=1}^{n} y_j$.

10600. *Proposed by Franz Rothe, University of North Carolina, Charlotte, NC.*
(a) Suppose a triangle has its vertices at integer lattice points in the plane and contains exactly 3 integer lattice points in its interior. Show the center of mass of the triangle is not an integer lattice point.
(b)* Find all values of i such that, if a triangle has its vertices at integer lattice points in the plane and contains exactly i integer lattice points in its interior, then the center of mass of the triangle cannot be an integer lattice point.

10601. *Proposed by Wen-Xiu Ma, Universität-GH Paderborn, Paderborn, Germany.* Let $n > 1$ be an integer and let $a_1, a_2, \dots, a_n$ be complex numbers. Show that

$$\begin{vmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{2n-1} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{2n-1} \\ \vdots & & & \ddots & \vdots \\ 1 & a_n & a_n^2 & \cdots & a_n^{2n-1} \\ 0 & 1 & 2a_1 & \cdots & (2n-1)a_1^{2n-2} \\ 0 & 1 & 2a_2 & \cdots & (2n-1)a_2^{2n-2} \\ \vdots & & & \ddots & \vdots \\ 0 & 1 & 2a_n & \cdots & (2n-1)a_n^{2n-2} \end{vmatrix} = (-1)^{n(n-1)/2} \prod_{1 \le i < j \le n} (a_i - a_j)^4.$$

10602. *Proposed by Dan Sachelarie, ICCE Bucharest, Romania, and Vlad Sachelarie, The Ohio State University, Columbus, OH.* In a triangle ABC, let H be the orthocenter, I the incenter, O the circumcenter, N the nine-point center, r the inradius, and R the circumradius. Prove that $\angle HIO \geq \pi/2 + \arcsin\sqrt{2r/R}$, with equality if and only if $\angle HIN = \pi/2$.

10603. *Proposed by Yury J. Ionin and Robin R. Lewis, Central Michigan University, Mt. Pleasant, MI.* Let a, b, and k be positive integers and let $P_k(a, b)$ be the period of the sequence $\{a^n \bmod b^k\}_{n=1}^{\infty}$. Find $\lim_{k\to\infty} P_{k+1}(a, b)/P_k(a, b)$.

10604. *Proposed by Joseph Rosenblatt, University of Illinois, Urbana, IL.*
(a) Determine positive constants c and C such that if $0 < a < b$ then

$$c\left(1 - \frac{a}{b}\right) \leq \sup_{x>0}\left|\frac{\sin(ax)}{ax} - \frac{\sin(bx)}{bx}\right| \leq C\left(1 - \frac{a}{b}\right). \qquad (*)$$

(b)* What are the largest constant c and the smallest constant C such that $(*)$ holds whenever $0 < a < b$?

10605. *Proposed by Jonathan M. Borwein and C. G. Pinner, Simon Fraser University, Burnaby, BC, Canada.* Let r and m be positive integers and define

$$P_r(m) = \prod_{n\neq m} \frac{n^r - m^r}{n^r + m^r}.$$

(a) Show that $P_1(m) = 0$ and that

$$P_3(m) = (-1)^{m+1}\frac{2}{3}(m!)^2 \prod_{n=1}^{m} \frac{n+m}{n^3 + m^3}.$$

(b) Show that $P_2(m) = (-1)^{m+1}\pi m/\sinh(\pi m)$ and that, more generally, $P_{2s}(m)$ is given by

$$(-1)^{m+1}\frac{2^{\epsilon} m\pi}{s}(\sinh m\pi)^{(-1)^s} \prod_{j=1}^{s-1}\left(\cosh\left(2\pi m \sin\left(\frac{j\pi}{2s}\right)\right) - \cos\left(2\pi m \cos\left(\frac{j\pi}{2s}\right)\right)\right)^{(-1)^j}$$

where $\epsilon = (1 + (-1)^s)/2$.

SOLUTIONS

Equilateral Cevian Triangles

10358* [1994, 76]. *Proposed by Jiang Huanxin, student, Fudan University, Shanghai, China.* In triangle ΔABC, find all points P such that the triangle ΔDEF (with $D = AP \cap BC$, $E = BP \cap CA$, $F = CP \cap AB$) is equilateral.

Solution by David Goering, Eastern Washington University, Cheney, WA. We restrict ourselves to the case where the points D, E, F belong to the line segments BC, AC, AB, respectively, in which case we say ΔDEF is *inscribed* in ΔABC.

The strategy is to characterize all equilateral triangles ΔDEF that can be inscribed in ΔABC, and then give a condition under which AD, BE, and CF are concurrent. Once this is done, we find P and prove its uniqueness.

All triangles up to similarity may be represented by the coordinates $A(0, 0)$, $B(1, 0)$, $C(b, c)$, with $c > 0$ and $0 \leq b \leq 1/2$. Let F be a point on AB with coordinates $F(x_0, 0)$. Then E on AC has coordinates

$$E\left(\frac{x_0 b \tan\phi}{b\tan\phi - c}, \frac{x_0 c \tan\phi}{b\tan\phi - c}\right),$$

where $\tan\phi$ is the slope of EF. If we now let M be the midpoint of EF and let s be the length of EF, we find the coordinates of D with the vector addition

$$D = M + \frac{\sqrt{3}}{2}s\langle\sin\phi, -\cos\phi\rangle.$$

This makes $\triangle DEF$ equilateral. We now have

$$D\left(\frac{x_0\big((2b+\sqrt{3}c)\tan\phi - c\big)}{2\big(b\tan\phi - c\big)}, \frac{cx_0\big(\tan\phi - \sqrt{3}\big)}{2\big(b\tan\phi - c\big)}\right).$$

Since we wish $D = (D_x, D_y)$ to lie on BC, for fixed ϕ, we find x_0 so that $D_y \cdot (b-1) = c \cdot (D_x - 1)$. Solving this equation for x_0 and substituting this value into the coordinates of D, E, and F gives coordinates for the equilateral triangle DEF. Letting $d = \sqrt{3} - \sqrt{3}b + c - (1 + b + \sqrt{3}c)\tan\phi$, we have

$$\begin{gathered}D\left(\frac{-(2b+\sqrt{3}c)\tan\phi + c}{d}, \frac{-c\big(\tan\phi - \sqrt{3}\big)}{d}\right),\\ E\left(\frac{-2b\tan\phi}{d}, \frac{-2c\tan\phi}{d}\right), \qquad F\left(\frac{2(c - b\tan\phi)}{d}, 0\right).\end{gathered} \tag{$*$}$$

The case $\phi = \pi/2$ is easily found by taking limits. Note that while every inscribed equilateral triangle is given by ($*$) for some value of ϕ, there are values of ϕ for which $\triangle DEF$ is *not* inscribed. That is, $\triangle DEF$ is an equilateral triangle whose vertices are on the extended lines AB, AC, and BC. Also, $\triangle DEF$ now has an orientation such that point D is to the right as FE is traversed from F to E. Finally, we note the geometric impossibility of an equilateral triangle with $0 \le \tan\phi \le \sqrt{3}$, since this forces segment FD to lie on or below AB.

We now express the equations of AD, BE, and CF in standard form and note that these lines are concurrent if and only if $\det \mathbf{M} = 0$, where $\mathbf{M}$ is the matrix of coefficients from the equations of the three lines. For the sake of brevity, we use the slope-intercept form, but the algebraic condition requires writing the slope of each of these lines in the form $(x + y\tan\phi)/(w + z\tan\phi)$ and clearing denominators. Denoting the slope of a line PQ by m_{PQ}, we have

$$\mathbf{M} = \begin{pmatrix} m_{AD} & -1 & 0 \\ m_{BE} & -1 & -m_{BE} \\ m_{CF} & -1 & c - bm_{CF} \end{pmatrix}.$$

After simplification, the equation $\det \mathbf{M} = 0$ reduces to

$$\begin{aligned}(-2b + 2b^2 - \sqrt{3}c + \sqrt{3}bc - 3c^2)\tan^3\phi + (5c - 13bc - \sqrt{3}c^2)\tan^2\phi \\ +(6b - 6b^2 - \sqrt{3}c + \sqrt{3}bc + 5c^2)\tan\phi + (-3c + 3bc - \sqrt{3}c^2) = 0\end{aligned}$$

Let $x = \tan\phi$ and let $p(x) = k_3x^3 + k_2x^2 + k_1x + k_0$ denote the above cubic in $\tan\phi$.

We now show that, for all admissible values of b and c, $p(x)$ has exactly one root that yields an inscribed $\triangle DEF$. Then P can be found as the intersection of any two of the lines AD, BE, and CF, and lies within $\triangle ABC$.

We begin by observing that $k_3 = 2b(b-1) + \sqrt{3}c(b-1) - 3c^2 < 0$, and $p(0) = 3c(b-1) - \sqrt{3}c^2 < 0$. This assures a negative real root.

Now consider the behavior of $p(x)$ at $x = -\sqrt{3}$. We have $p(-\sqrt{3}) = 24c(1-2b) \ge 0$, which requires that there be at least one real root in $\big[-\sqrt{3}, 0\big)$. If $\triangle ABC$ is isosceles ($b = 1/2$), then $p(-\sqrt{3}) = 0$. When $\tan\phi = -\sqrt{3}$, $F_x = 1/2$ and $m_{ED} = 0$, as we would expect. If $b < 1/2$, $p(-\sqrt{3}) > 0$. We also have $p'(-\sqrt{3}) = -16c^2 + 4\sqrt{3}c(9b - 5) + 12b(b-1) < 0$, and $p''(-\sqrt{3}) = -12\sqrt{3}b^2 - 44bc + 16\sqrt{3}c^2 + 12\sqrt{3}b + 28c =$

$12\sqrt{3}b(1-b)+4c(7-11b)+16\sqrt{3}c^2>0$, for $c>0$ and $0\le b\le 1/2$. Since $p(x)$ is nonnegative, decreasing, and concave up at $x=-\sqrt{3}$, $p(x)$ has no roots in $\left(-\infty,-\sqrt{3}\right)$. Thus all negative roots of p are in $\left[-\sqrt{3},0\right)$.

In a similar manner, it can be shown that, for $x=\sqrt{3}$, $p(x)$ is negative, decreasing and concave down. Thus, $p(x)$ has no roots in $\left[\sqrt{3},\infty\right)$. Since $\triangle DEF$ cannot be inscribed for $0\le\tan\phi<\sqrt{3}$, $p(x)$ has no roots in $[0,\infty)$ that yield an inscribed triangle.

To show that $p(x)$ cannot have more than one root in $\left[-\sqrt{3},0\right)$, we suppose that it does and argue by contradiction. Since $p(0)<0$, $p(x)$ must have a local maximum in $\left(-\sqrt{3},0\right)$, and it must be the case that $p'(0)<0$ and $p''(0)<0$. Since $p'(0)=k_1$ and $p''(0)=2k_2$, we consider the graphs of $k_1<0$ and $2k_2<0$. The former region is bounded by a hyperbola in b and c; the latter by two intersecting lines. From their graphs, it is clear that $p'(0)$ and $p''(0)$ are never simultaneously negative in $[0,1/2)\times(0,\infty)$, the geometrically relevant region. Thus, $p(x)$ has exactly one root in $\left[-\sqrt{3},0\right)$.

We now show that the equilateral triangle determined by setting $\tan\phi$ equal to the negative root of $p(x)$ is inscribed in $\triangle ABC$, i.e., that D, E, and F defined by $(*)$ belong to the respective sides of $\triangle ABC$, rather than simply to their extensions. First, note that $-\sqrt{3}\le\tan\phi<0$ implies that the coordinates of D, E, and F are all nonnegative ($d=\sqrt{3}(1-b)+c-(1+b+\sqrt{3}c)\tan\phi>0$). It can be shown algebraically that $E_x<b$, which requires point E to lie on the open segment AC. It can also be shown algebraically that $F_x\ge 1$ implies that $D_x>b$. The remaining possibilities are 1) D and E belong to $\triangle ABC$ while $F_x\ge 1$, and 2) E and F belong to $\triangle ABC$ while $D_x\le b$.

Suppose that $F_x>1$. Then the point $AD\cap BE$ lies in the interior of $\triangle ABC$, but $CF\cap BE$ lies outside $\triangle ABC$, so AD, BE, and CF are not concurrent.. But this contradicts the fact that $\tan\phi$ is a solution to $\det\mathbf{M}=0$. Thus, $F_x<1$. Similarly, if $D_x<b$, $BE\cap FC$ is inside $\triangle ABC$, but $BE\cap AD$ is not, so again the three lines are not concurrent, which is a contradiction. This establishes that points D, E, and F lie on the open segments BC, AC, and AB, respectively, when $\tan\phi$ is given by the negative root of $p(x)$.

We have thus established that $p(x)$ has exactly one root in $\left[-\sqrt{3},0\right)$ and that this root determines an equilateral triangle inscribed in $\triangle ABC$. Furthermore, the lines AD, BE, and CF are concurrent and meet inside $\triangle ABC$ at point P. For $0\le b\le 1/2$ and $c>0$, we have

$$P_x=\frac{2\tan\phi\big((2b+\sqrt{3}c)\tan\phi-c\big)}{3-3b+\sqrt{3}c+2(-\sqrt{3}+\sqrt{3}b-3c)\tan\phi+(1+3b+3\sqrt{3}c)\tan^2\phi}$$

and

$$P_y=P_x\frac{c\left(\tan\phi-\sqrt{3}\right)}{(2b+\sqrt{3}c)\tan\phi-c}$$

where $\tan\phi$ is the unique negative real root of $p(x)$.

Editorial comment. Two partial solutions were received. Charles H. Jones established the existence of a suitable choice of P as follows. For each D in the segment BC, choose P in the segment AD and determine E and F as in the statement of the problem. There is a unique choice of P for which $\angle EDF=\pi/3$. The curve of such points $P(D)$ goes from B to C as D goes from B to C, and $|DF|-|DE|$ changes sign on this curve. By the intermediate value theorem, there exists a choice with $|DF|=|DE|$. He also studied the existence of solutions when $\triangle DEF$ is not required to be inscribed in the sense of the selected solution. If the point P is on any of the extended sides of $\triangle ABC$, or on the line parallel to one of the sides through the opposite vertex, then the $\triangle DEF$ is never equilateral. Removing these six lines leaves 16 connected regions that can be studied by analytic means. Some conditions for the existence of solutions in each of these regions were obtained.

Helen M. Marston considered only inscribed equilateral triangles, and suggested using Ceva's theorem to arrive at an equation characterizing the point P.

A Recurrence Whose Solution Changes Sign Infinitely Often

10396 [1994, 681]. *Proposed by Ioan Tomescu, University of Bucharest, Bucharest, Romania.* Let $\alpha > 0$ and let $\langle b_n : n \geq 1\rangle$ be defined recursively by $b_1 = \alpha$, $b_2 = 3\alpha$, and

$$b_{n+1} = (2n+1)b_n - (n^2 + \alpha^2)b_{n-1} \quad (n \geq 2).$$

Prove that $\langle b_n\rangle$ contains infinitely many positive and infinitely many negative terms.

Solution I by Richard Stong, Rice University, Houston, TX. We show that b_n is the imaginary part of c_n, where $c_n = (1+i\alpha)(2+i\alpha)\cdots(n+i\alpha)$. First note that c_n satisfies the given recurrence. To see this, write

$$\begin{aligned}\frac{c_{n+1}}{c_{n-1}} &= (n+1+i\alpha)(n+i\alpha) = (2n+1)(n+i\alpha) + (-n+i\alpha)(n+i\alpha)\\ &= (2n+1)\frac{c_n}{c_{n-1}} - (n^2+\alpha^2).\end{aligned}$$

Since the recurrence is linear with real coefficients, b_n is also a solution. One then easily checks that it has the desired initial values.

The required property of b_n is now clear. The argument of c_n is $\sum_{k=1}^{n} \arctan(\alpha/k)$. This sum diverges like the harmonic series since all terms are between 0 and $\pi/2$ and, for large k, the summand is asymptotic to α/k. Therefore, c_n must be in each quadrant of the complex plane infinitely often and, in particular, b_n must take on both signs infinitely often.

Solution II by Richard Holzsager, American University, Washington, DC. Let $q_n = b_n/b_{n-1}$. The recursion is then $q_{n+1} = 2n+1 - (n^2+\alpha^2)/q_n$. If $q_n < 0$, then b_n and b_{n-1} have opposite signs; while if q_n is 0 or ∞, then $b_k = 0$ for $k = n$ or $n-1$, so the original recurrence shows that b_{k+1} and b_{k-1} have opposite signs. Thus it suffices to show that there are arbitrarily large n for which q_n is not a positive real number. It is convenient to write $d_n = q_n - n$. Then

$$d_{n+1} = n - \frac{n^2+\alpha^2}{n+d_n} = \frac{nd_n - \alpha^2}{n+d_n} = d_n - \frac{\alpha^2 + d_n^2}{n+d_n}, \qquad (*)$$

and we need to show that assuming $d_n > -n$ for all $n \geq N$ leads to a contradiction. With this hypothesis, $(*)$ implies that $d_N > d_{N+1} > d_{N+2} > \cdots$. But then, for $n > N$, $(\alpha^2 + d_n^2)/(n+d_n) > \alpha^2/(n+d_N)$. Since $\sum \alpha^2/(n+d_N)$ is divergent, it follows that d_n diverges to $-\infty$.

For $0 > d_n > -n$, $|d_{n+1}| \geq |d_n|\left(1 + |d_n|/n\right)$, so

$$\frac{|d_{n+1}|}{n+1} \geq \frac{|d_n|}{n}\cdot\frac{1+|d_n|/n}{1+1/n}.$$

Since $|d_n|$ is eventually greater than 2, one has for all large n

$$\frac{1+|d_n|/n}{1+1/n} > \frac{1+2/n}{1+1/n}.$$

These form a divergent product, so the d_n/n must also diverge to $-\infty$, contradicting the assumption that $d_n > -n$.

Solution III by Donald A. Darling, Newport Beach, CA. Define a sequence $\{a_n\}, n = 1, 2, \ldots$ by setting $b_n = n!a_n$. Let $f(x) = \sum_{n=1}^{\infty} a_n x^n$ be the generating function for the sequence $\{a_n\}$. We show that

$$f(x) = \frac{1}{1-x}\sin\left(\alpha \log\frac{1}{1-x}\right), \qquad |x| < 1. \qquad (*)$$

This function has infinitely many zeros in the interval $0 \leq x < 1$, namely at the points $x_j = 1 - \exp(-2j\pi/\alpha)$, $j = 0, 1, 2, \ldots$, so the extension of Descartes rule of signs to power series shows that a_n, and hence b_n, has infinitely many changes in sign. This extension of Descartes rule of signs may be found in G. Pólya and G. Szegő, *Problems and Theorems in Analysis*, Vol. II, Springer-Verlag, 1972, Part 5, Chapter 1.

To prove $(*)$, we first set $b_n = n!a_n$ in the difference equation for $\{b_n\}$, and after some routine simplification, we obtain

$$a_{n+1} - 2a_n + a_{n-1} = -\frac{1}{n+1}(a_n - a_{n-1}) - \frac{\alpha^2}{n(n+1)}a_{n-1}, \qquad (n \geq 2) \qquad (**)$$

with $a_1 = \alpha$ and $a_2 = 3\alpha/2$. Multiplying the recurrence $(**)$ by $n(n+1)x^{n-1}$, summing for $n \geq 2$, and using the initial values for a_1 and a_2 yields the differential equation

$$(1-x)^2 f''(x) - 3(1-x)f'(x) + (1+\alpha^2)f(x) = 0$$

through straightforward calculation. This is a classical Euler differential equation in the variable $(1-x)$, which can be transformed into an equation with constant coefficients. The solution with the initial conditions $f(0) = 0$, $f'(0) = a_1 = \alpha$, is given by $(*)$.

Editorial comment. Several solvers observed that all nonzero solutions of the recurrence have the required property. In particular, Leandro Cagliero obtained this result using the method of Solution II. He also remarked that the solution of the recurrence with $\alpha = 0$, $b_1 = A$, and $b_2 = B$ is $b_n = \big((3A - B) + (B - 2A)\sum_{j=1}^{n}(1/j)\big)n!$, which has constant sign for all sufficiently large n.

The solution of Th. B. van Dulken was based on a comparison theorem for difference equations analogous to those for ordinary differential equations. This theorem allowed him to obtain some general results on the oscillatory behavior of solutions of difference equations.

Solved also by J. Alvarez (Spain), J. Anglesio (France), D. Borwein (Canada), L. Cagliero (Argentina), Th. B. van Dulken (Australia), P. G. Kirmser, J. H. Lindsey II, S. C. Locke, L. E. Mattics, I. Nemes (Austria), A. Nijenhuis, V. Novakov (Bulgaria), M. Vowe (Switzerland), A. N. 't Woord (The Netherlands), WMC Problems Group, and the proposer.

Asymptotic Behavior of a Nonexpansive Sequence

10404 [1994, 792]. *Proposed by Behzad Djafari Rouhani, Shahid Beheshti University and Islamic Azad University, Tehran, Iran.* Let $x_1, x_2, \ldots$ be a sequence of real numbers such that $|x_i - x_j| \geq |x_{i+1} - x_{j+1}|$ for all positive integers i, j with $|i - j| \leq 2$. Prove that $\langle x_n/n \rangle$ converges to a finite limit as $n \to \infty$.

Solution by Allen Stenger, Irvine, CA. The sequences $\{\,|x_n - x_{n+1}|\,\}$ and $\{\,|x_n - x_{n+2}|\,\}$ are nonincreasing and bounded below by zero, so they must approach limits. If either limit is zero, i.e., $x_n - x_{n+1} = o(1)$ or $x_n - x_{n+2} = o(1)$, then by telescoping summation we infer that $x_n = o(n)$, hence $\{x_n/n\}$ approaches zero.

Therefore assume neither limit is zero. We prove the stronger statement that $\{x_n - x_{n+1}\}$ approaches a finite limit, from which we again infer by summation that $\{x_n/n\}$ approaches a limit.

Denote the limit of $\{\,|x_n - x_{n+1}|\,\}$ by L (where $L > 0$), and write $x_n - x_{n+1} = Ls_n + o(1)$, where $s_n = \pm 1$. Then

$$|x_n - x_{n+2}| = |\,(x_n - x_{n+1}) + (x_{n+1} - x_{n+2})\,| = L\,|s_n + s_{n+1}| + o(1).$$

Since this approaches a nonzero limit, we infer that s_n must have a constant sign after some point and is therefore a constant s. Thus, $x_n - x_{n+1} = Ls + o(1)$ as claimed.

Editorial comment. The proposer noted that the problem arose from a study of the theorem of A. Pazy, Asymptotic behavior of contractions in Hilbert space, *Israel J. Math.* 9 (1971), 235–240.

Solved also by R. Barbara (Lebanon), D. Borwein (Canada), P. Budney, R. J. Chapman (U. K.), D. Cui, J.-P. Grivaux (France), R. Holzsager, G. Keselman & R. R. Goldberg, J. H. Lindsey II, O. P. Lossers (The Netherlands), H. Morris, A. Nijenhuis, C. G. Petalas & T. P. Vidalis (Greece), C. Popescu (Belgium), F. Richman, K. Schilling, M. Shemesh (Israel), R. Stong, A. A. Tarabay (Lebanon), NSA Problems Group, Prague Problem Solution Group (Czech Republic), USA Problems Group, WMC Problems Group, and the proposer.

Brianchon, Desargues, Pascal

10405 [1994, 793]. *Proposed by Herbert Gülicher, Westfälische Wilhelms-Universität, Münster, Germany.* Let $A_1A_2A_3A_4A_5A_6$ be a hexagon *circumscribed* about a conic, and form the intersections $P_i = A_iA_{i+2} \cap A_{i+1}A_{i+3}$ ($i = 1, \ldots, 6$, all indices mod 6). Show that the P_i are the vertices of a hexagon *inscribed* in a conic.

Solution by Albert Nijenhuis, University of Pennsylvania (Emeritus), Philadelphia, PA, and University of Washington, Seattle, WA. Since $A_1A_2A_3A_4A_5A_6$ is circumscribed about a conic, by Brianchon's Theorem the lines A_iA_{i+3} ($i = 1, 2, 3$) are concurrent. Let $\mathcal{O}$ be the common point.

By Desargues's Theorem, since the triangles $A_iA_{i+2}A_{i+4}$ ($i = 1, 2$) are perspective (with respect to $\mathcal{O}$), the intersections of the pairs of corresponding sides $A_3A_5 \cap A_6A_2$, $A_1A_5 \cap A_4A_2$, and $A_1A_3 \cap A_4A_6$ are collinear.

By (the converse of) Pascal's Theorem, since the pairs (P_1P_2, P_4P_5), (P_2P_3, P_5P_6), and (P_3P_4, P_6P_1) intersect in collinear points, the points $P_1, \ldots, P_6$ lie on a conic.

Editorial comment. R. H. Jeurissen remarked that each of the three classical theorems used here gives an equivalence between two conditions, so that the argument can be reversed to show that $A_1A_2A_3A_4A_5A_6$ is circumscribed about a conic if $P_1, \ldots, P_6$ lie on a conic. In fact, this converse is the projective dual of the present problem, since the construction of sides of $A_1A_2A_3A_4A_5A_6$ from the sides of $P_1P_2P_3P_4P_5P_6$ is dual to the given construction. Even the proof is self-dual since Desargues's Theorem is self-dual and the other two are duals of each other. The proposal was accompanied by a proof by Günther Pickert using properties of the Steiner points of an inscribed hexagon in place Desargues's Theorem. Further details about these classical theorems of projective geometry can be found in H. S. M. Coxeter, *Introduction to Geometry*, second edition, Wiley, 1969, Heinrich Dörrie, *100 Great Problems of Elementary Mathematics*, Dover, 1965, and D. Pedoe, *A Course of Geometry*, Cambridge, 1970.

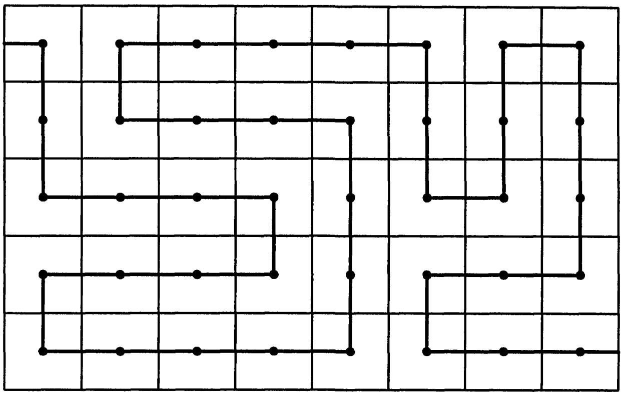
Solved also by J. Anglesio (France), M. Benedicty, R. H. Jeurissen (The Netherlands), O. P. Lossers (The Netherlands), G. Pickert, C. G. Petalas & T. P. Vidalis (Greece), C. Popescu (Belgium), M. Reid, R. Tauraso (Italy), and the proposer.

Grid Paths Split Area in Half

10406 [1994, 793]. *Proposed by David C. Fisher, University of Colorado, Denver, CO, Karen L. Collins, Wesleyan University, Middleton, CT, and Lucia B. Krompart, Rochester, MI.* Suppose a path on an m-by-n square grid starts at the northwest corner, goes through each point exactly once, and ends at the southeast corner. Show that such a path divides the grid into two equal halves: (a) those regions opening north or east, and (b) those regions opening south or west.

Solution I by Roger B. Eggleton, Illinois State University, Normal, IL, and R.J. Simpson, Curtin University of Technology, Perth, Western Australia. Embed the $m \times n$ grid in an $m \times n$ array of unit squares centered at the grid points. Extend each end of the curve by a half-unit horizontal segment outward to the edge of the array. (See the figure at the top of the next page.) Including the squares adds congruent L-shaped areas to the two regions described, so it suffices to show that the augmented regions A and B have equal area.

In traversing the curve, region A is to the left and region B is to the right. When the curve turns right in a cell, it adds $3/4$ to the area of A and $1/4$ to the area of B. When it turns left, the contributions are $1/4$ and $3/4$, respectively. Cells with no turn are split in half. Since



the curve does not intersect itself and has parallel initial and final segments, the number of left turns equals the number of right turns. Thus the areas are equal.

Solution II by Michael Reid, Brown University, Providence, RI. By Pick's Theorem, the area of a polygon with vertices at lattice points depends only on the number i of interior lattice points and the number b of boundary lattice points. (It equals $i + b/2 - 1$, but we don't need this.)

Complete the path between $(1, n)$ and $(m, 1)$ to a polygon by adding segments from the ends to $(0, 0)$. Alternatively, add segments from the ends to $(m + 1, n + 1)$. The polygons augment the desired regions by congruent figures, so it suffices to show that they have the same area. Each has $mn + 1$ boundary points and no interior points, so by Pick's Theorem the areas are equal.

Editorial comment. Most solvers used Pick's Theorem, which is discussed in Grünbaum and Shepard, Pick's Theorem, this MONTHLY **100** (1993), 150–161. Stephen H. Schanuel observed that the argument using Pick's Theorem applies in the more general problem where the segments of the path are not constrained to the grid lines.

The desired result also follows from Grinberg's Theorem in graph theory. Adding a curve from $(1, n)$ to $(m, 1)$ outside the grid completes a spanning cycle in a planar graph with mn vertices. The unbounded face and the long face inside the cycle have length $m + n - 1$. The remaining faces are 4-cycles. Grinberg's Theorem states that (length$(F) - 2$) has the same sum over the faces F inside the cycle as it does over the faces outside, and therefore there are as many squares inside as outside.

Solved also by J. C. Barthelmes, D. Beckwith, M. Benedicty, S. Brandt (Germany), L. Cagliero & J. Lauret (Argentina), R. J. Chapman (U. K.), R. Ehrenborg (Canada), J. W. Grossman, R. Holzsager, R. H. Jeurissen & B. Polman (The Netherlands), U. Klein (Germany), N. Komanda, J. Kupka (Australia) & M. Leinert (Germany), K. M. Levasseur, J. H. Lindsey II, O. P. Lossers (The Netherlands), D. Marcus, R. F. McCoart, A. Nijenhuis, C. Popescu (Belgium), R. C. Read (Canada), R. M. Robinson, S. H. Schanuel, R. Stong, D. Wolfe, Anchorage Problem Solutions Group, NSA Problems Group, Prague Problem Solution Group (Czech Republic), WMC Problems Group, and the proposers.

The Effect of Truncation

10419 [1994, 1014]. *Proposed by Bill Correll, Jr., Denison University, Granville, OH.* Let k be an integer greater than or equal to 3. Let $S(k)$ be the set of nonnegative real numbers x for which

$$\left\lfloor \frac{x+k-2}{k} \right\rfloor \left\lfloor \frac{x+k-1}{k-1} \right\rfloor + \left\lfloor \frac{x}{k} \right\rfloor = \left\lfloor \frac{x+k-2}{k-1} \right\rfloor \left\lfloor \frac{x+k-1}{k} \right\rfloor + \left\lfloor \frac{x}{k-1} \right\rfloor .$$

(**a**) Determine the largest integer in $S(k)$.

(**b**) Show that $S(k)$ is the union of a finite number of intervals with the sum of the lengths of those intervals equal to $(k^2 - 3k + 6)/2$.

Composite solution by O. P. Lossers, Technical University Eindhoven, Eindhoven, The Netherlands, and National Security Agency Problems Group, Fort Meade, MD. The largest integer in $S(k)$ is $k(k-2)$. Since the equation holds for x if and only if it holds for $\lfloor x \rfloor$, it suffices to determine the integers in $S(k)$.

The desired equation is $AD + \lfloor x/k \rfloor = BC + \lfloor x/k - 1 \rfloor$, where $A = \lfloor (x+k-2)/k \rfloor$, $B = \lfloor (x+k-1)/k \rfloor$, $C = \lfloor (x+k-2)/(k-1) \rfloor$, and $D = \lfloor (x+k-1)/(k-1) \rfloor$. When x is an integer, we have $A = B - \epsilon$ and $C = D - \epsilon'$, where ϵ is 1 if $x \equiv 1 \bmod k$ and 0 otherwise, and ϵ' is 1 if $x \equiv 0 \bmod (k-1)$ and 0 otherwise.

When $A = B$ and $C = D$, we require $\lfloor x/k \rfloor = \lfloor x/(k-1) \rfloor$. Letting $x = a(k-1) + b$, the equation holds when $0 \le a \le b < k-1$. We must exclude $b \in \{0, a+1\}$. There remain $\sum_{b=1}^{k-2} b = (k-1)(k-2)/2$ solutions, of which $x = (k-2)k$ is the largest.

When $A = B - 1$ and $C = D$, we require $\lfloor x/k \rfloor = \lfloor x/(k-1) \rfloor + D > \lfloor x/(k-1) \rfloor$. This inequality fails for $x \ge 0$ and $k \ge 3$, so there are no solutions in this case.

When $A = B$ and $C = D - 1$, we require $\lfloor x/k \rfloor + \lfloor (x+k-1)/k \rfloor = \lfloor x/(k-1) \rfloor$. Also x must be divisible by $k-1$ and not congruent to 1 modulo k. This holds only when x is 0 or $k-1$.

When $A = B - 1$ and $C = D - 1$, we require $\lfloor x/k \rfloor + \lfloor (x+k-1)/k \rfloor = \lfloor x/(k-1) \rfloor + \lfloor (x+k-1)/(k-1) \rfloor$. Also $x \equiv 0 \bmod (k-1)$ and $x \equiv 1 \bmod k$. Since x is a multiple of $k-1$, we have $\lfloor (x+k-1)/(k-1) \rfloor > \lfloor (x+k-1)/k \rfloor$, and always $\lfloor x/(k-1) \rfloor \ge \lfloor x/k \rfloor$. Thus equality cannot hold in this case.

Altogether, we obtain $(k-1)(k-2)/2 + 2 = (k^2 - 3k + 6)/2$ unit-length half-open intervals where the condition holds, with the largest integer being $k(k-2)$.

Solved also by J. C. Binz (Switzerland), R. J. Chapman (U. K.), J. Christopher, J. H. Lindsey II, D. K. Nester, A. A. Tarabay (Lebanon), D. B. Tyler, A. N. 't Woord (Netherlands), WMC Problems Group, and the proposer.

Sums of Two Squares and a Cube

10426 [1995, 70]. *Proposed by Noam Elkies, Harvard University, Cambridge, MA, and Irving Kaplansky, Mathematical Sciences Research Institute, Berkeley, CA.* Show that any integer can be expressed as a sum of two squares and a cube. Note that the integer being represented and the cube are both allowed to be negative.

Solution by Andrew Adler, University of British Columbia, Victoria, British Columbia, Canada.

$$\begin{aligned}
2x + 1 &= (x^3 - 3x^2 + x)^2 + (x^2 - x - 1)^2 - (x^2 - 2x)^3 \\
4x + 2 &= (2x^3 - 2x^2 - x)^2 + (2x^3 - 4x^2 - x + 1)^2 - (2x^2 - 2x - 1)^3 \\
8x + 4 &= (x^3 + x + 2)^2 + (x^2 - 2x - 1)^2 - (x^2 + 1)^3 \\
16x + 8 &= (2x^3 - 8x^2 + 4x + 2)^2 + (2x^3 - 4x^2 - 2)^2 - (2x^2 - 4x)^3 \\
16x &= (x^3 + 7x - 2)^2 + (x^2 + 2x + 11)^2 - (x^2 + 5)^3
\end{aligned}$$

Editorial comment. Other identities were supplied by readers, but all solutions used a similar division into cases. John P. Robertson notes that the representation for odd integers follows from Theorem 2 on page 113 of L. J. Mordell, *Diophantine Equations*, Academic Press, 1966. He and the proposers mention related open problems, including sums of a square and two cubes.

Solved also by J. P. Robertson, A. N. 't Woord (The Netherlands), and the proposers.

A Polynomial Identity

10466 [1995, 654]. *Proposed by E. Sparre Andersen & Mogens Esrom Larsen, University of Copenhagen, Copenhagen, Denmark.* For $x \in \mathbb{C}$ and $n \in \mathbb{N}$, prove the following identities between polynomials.

$$\textbf{(a)} \qquad (-4)^n \sum_{j=0}^{n} \binom{x+1/2}{j}\binom{n-1-x}{2n-j} = \binom{2n}{n} \sum_{j=0}^{n} \binom{x+j}{2j}\binom{x-j}{2n-2j}.$$

For all $m \in \mathbb{N}$ with $0 \le m \le 2n$, generalize (**a**) to

$$\textbf{(b)} \quad (-4)^n \sum_{j=0}^{n} \binom{x+1/2}{j}\binom{n-1-x}{2n-j} = \binom{2n}{n} \sum_{j=-\lfloor m/2 \rfloor}^{n-\lceil m/2 \rceil} \binom{x+j}{2j+m}\binom{x-j}{2n-m-2j}.$$

Solution by Robin J. Chapman, University of Exeter, Exeter, UK. Since (**b**) reduces to (**a**) when $m = 0$, it suffices to prove (**b**). Both sides are polynomials in x of degree at most $2n$, so it suffices to verify that they are equal for at least $2n+1$ distinct values of x. We use the n integers and the $n+1$ half-integers in the interval $[-1/2, n-1/2]$. We show that both sides equal 0 at the integers and equal $4^{-n}\binom{2n}{n}$ at the half-integers in this range. Let

$$f(x) = \sum_{j=0}^{n} \binom{x+1/2}{j}\binom{n-1-x}{2n-j} \quad \text{and} \quad g_m(x) = \sum_{j=-\lfloor m/2 \rfloor}^{n-\lceil m/2 \rceil} \binom{x+j}{m+2j}\binom{x-j}{2n-m-2j}.$$

First consider f. When $0 \le x \le n-1$ and $0 \le j \le n$, we have $2n-j > n-1-x$. When x is an integer in this range, we thus have $\binom{n-1-x}{2n-j} = 0$ and $f(x) = 0$. When $x = r - 1/2$ with r a nonnegative integer, the sum for $f(x)$ reduces to $\sum_{j=0}^{r} \binom{r}{j}\binom{n-r-1/2}{2n-j}$. Since the Vandermonde convolution $\binom{y}{s} = \sum_{j=0}^{r} \binom{r}{j}\binom{y-r}{s-j}$ is valid for real y and nonnegative integer r, we have $f(r-1/2) = \binom{n-1/2}{2n} = (-16)^{-n}\binom{2n}{n}$, as desired.

We next show that $g_m(x) = 0$ when x is an integer in $[0, n-1]$. If $g_m(x) \neq 0$, then some summand $\binom{x+j}{m+2j}\binom{x-j}{2n-m-2j}$ must be nonzero. Both factors must be nonzero; the first requires a) $x+j < 0$ or b) $x+j \ge m+2j$, and the second requires c) $x-j < 0$ or d) $x-j \ge 2n-m-2j$. Each choice produces a contradiction: (a&c) $\implies$ $2x < 0$; (b&d) $\implies$ $2x \ge 2n$; (a&d) $\implies$ $2n-m < 0$; (b&c) $\implies$ $m < 0$. Thus each term is 0.

It remains to evaluate $g_m(x)$ at half-integer values. We first subtract 0 in the form

$$h_m(x) = \sum_{j=-\lfloor (m+1)/2 \rfloor}^{n-\lceil (m+1)/2 \rceil} \binom{x+j+1/2}{m+2j+1}\binom{x-j-1/2}{2n-m-2j-1}$$

The summand for $h_m(x)$ is that for $g_m(x)$ with $j+1/2$ substituted for j. Since $x+j+1/2$ is now integral, the same case analysis with $j+1/2$ in place of j shows that every term in the sum is 0 except possibly when $x+j+1/2 < 0$ and $x-j-1/2 < 0$. This requires $x = -1/2$, $j < 0$, and $j > -1$, which again is impossible.

We now let $k_m(x) = g_m(x) - h_m(x)$. By setting $s = m+2j$ in the summand for $g_m(x)$ and $s = m + 2(j+1/2)$ in the summand for $h_m(x)$ and observing that $m-s$ is even for terms in the first sum and odd for terms in the second, we obtain

$$k_m(x) = (-1)^m \sum_{s=0}^{2n} (-1)^s \binom{x - \frac{m-s}{2}}{s}\binom{x + \frac{m-s}{2}}{2n-s}.$$

We claim that $k_m(x) = 4^{-n}$ for all x. Rearranging the factors in computing binomial coefficients yields

$$\binom{x - \frac{m-s}{2}}{s}\binom{x + \frac{m-s}{2}}{2n-s} = \frac{m!(2n-m)!}{s!(2n-s)!}\binom{x + \frac{m-s}{2}}{m}\binom{x - \frac{m-s}{2}}{2n-m}.$$

Define $\phi(z) = \binom{(z+m)/2}{m}$ and $\psi(z) = \binom{(z-m)/2}{2n-m}$. With $y = 2x$, we obtain

$$k_m(x) = (-1)^m \binom{2n}{m}^{-1} \sum_{s=0}^{2n} (-1)^s \binom{2n}{s} \phi(y-s)\psi(y+s).$$

From the calculus of finite differences, it follows that the sum $\sum_{s=0}^{2n}(-1)^s\binom{2n}{s}s^p$ is 0 when p is an integer with $0 \le p < 2n$ and is $(2n)!$ when $p = 2n$. For nonnegative integers p, q, the expression $(y-s)^p(y+s)^q$ is a homogeneous polynomial of degree $p+q$. Thus the sum

$$\sum_{s=0}^{2n}(-1)^s\binom{2n}{s}(y-s)^p(y+s)^q$$

is 0 when $p+q < 2n$ and equals $(-1)^p(2n)!$ when $p+q = 2n$.

Since $\phi(z)$ and $\psi(z)$ are polynomials in z, we can express the summand in $k_m(x)$ as a linear combination of terms of the form $(y-s)^p(y+s)^q$. Since the leading terms of $\phi(z)$ and $\psi(z)$ are $z^m/(2^m m!)$ and $z^{2n-m}/(2^{2n-m}(2n-m)!)$, respectively, the only nonzero contribution is

$$k_m(x) = (-1)^m \binom{2n}{m}^{-1} \frac{(-1)^m(2n)!}{2^m m! 2^{2n-m}(2n-m)!} = 4^{-n}.$$

Solved also by S. B. Ekhad, C. Krattenthaler (Austria), Con Amore Problem Group (Denmark), and the proposers.

Solutions Need Not Tend to Zero

10498 [1996, 75]. *Proposed by Ray Redheffer, University of California, Los Angeles, CA.* Consider the system of differential equations

$$\frac{dx}{dt} = -\big(x + a(t)y\big) \qquad \frac{dy}{dt} = -\big(b(t)x + y\big) \qquad (*)$$

where $a(t)$ and $b(t)$ are positive, continuous, and bounded for $0 \le t < \infty$.

If $\big(\sup a(t)\big)\big(\sup b(t)\big) < 1$, it is easy to prove that all solutions of $(*)$ tend to 0 as $t \to \infty$. Does the same conclusion follow if one assumes only that $\sup\big(a(t)b(t)\big) < 1$?

Solution by the University of South Alabama Problem Group, Mobile, AL. We give an example to show that the answer is no. Take $a(t)$ to be periodic with period 1 and define

$$a(t) = \begin{cases} 36 & \text{if } 0 \le t < 1/4; \\ 36 - (1439/5)(t-1/4) & \text{if } 1/4 \le t < 3/8; \\ 1/40 & \text{if } 3/8 \le t < 7/8; \\ 1/40 + (1439/5)(t-7/8) & \text{if } 7/8 \le t < 1. \end{cases}$$

Now set $b(t) = a(t-1/2)$ and compute $\sup\big(a(t)b(t)\big) = 9/10$. Take $x(0) = 1$ and $y(0) = -1$ for the initial conditions and set $u(t) = e^t x(t)$ and $v(t) = -e^t y(t)$. Then the initial value problem becomes:

$$\frac{du}{dt} = a(t)v, \quad \frac{dv}{dt} = b(t)u, \quad u(0) = 1, \quad v(0) = 1.$$

Clearly $u(t)$ and $v(t)$ are increasing functions. Using the mean value theorem we have $u(n+1/4) - u(n) > 36v(n)\cdot 1/4$ and $v(n+3/4) - v(n+1/2) > 36u(n+1/4)\cdot 1/4$ for $n = 0, 1, 2, \ldots$. Hence $u(n+1) - u(n) > 9v(n)$ and $v(n+1) - v(n) > 9u(n)$ for $n = 0, 1, 2, \ldots$, so we get $u(n) > 10^n$ and $v(n) > 10^n$ by induction. Hence $x(n) > e^{-n}10^n$, $y(n) < -e^{-n}10^n$, and we have our example.

Solved also by P. Alsholm (Denmark), J. C. Bronski, R. Kelsey, J. H. Lindsey II, L. Scribani (South Africa), and the proposer.

REVIEWS

Edited by **Underwood Dudley**
Mathematics Department, De Pauw University, Greencastle, IN 46135

The Life of Stefan Banach. By Roman Kałuża. Translated and edited by Ann Kostant and Wojbor Wocyzyński. Birkhäuser, Boston, 1996, x + 137, $24.50.

Reviewed by **Sheldon Axler**

In an least one printing of the current (fifteenth) edition of the *Encyclopedia Britannica*, the entry on Stefan Banach did not contain the words "Poland" or "Polish". The *Britannica* called Banach a "Soviet mathematician." The encyclopedia fixed its error in later printings, but the mathematics community has not yet adequately documented Banach's life and ideas. A computer search of *Mathematical Reviews* reveals more than eleven thousand publications with the word "Banach" in the title; "Hilbert" occurs in only seven thousand titles. Yet no mathematician or historian of mathematics has produced a book-length biography of Stefan Banach.

The book under review was written neither by a mathematician nor by a historian. The author, a Polish reporter and journalist, writes well about mathematics without using any mathematical symbols. Professional mathematicians will spot a few technical errors of the type that inevitably creep into exposition at this level. For example, we read that "the only linear transformations" on a finite-dimensional Euclidean space are "translations, rotations, and reflections." Such small mistakes in mathematical details can easily be forgiven because the author does a good job of capturing the flavor of early functional analysis and its creators.

The book suffers more from the lack of a historian's perspective than from an absence of mathematical expertise. Some events described in the book cry out for more explanation. For example, consider the author's description of the Nazi efforts to eliminate the intelligentsia in occupied Poland during World War II. Before capturing the Polish university town of Lvov, where Banach lived and worked, German officials compiled a list of prominent professors, scientists, and writers in Lvov who would be executed. One night shortly after German soldiers had entered Lvov, SS units murdered forty leading intellectual figures in Lvov without even the pretense of trials. But Banach was untouched by the Nazi death squads. An alert reader will wonder why Banach, who at this time was President of the Polish Mathematical Society and a Dean at the university, was not among the intellectuals marked down for liquidation. Unfortunately the author does not comment on the apparent disparity between his description of Nazi plans to crush Polish intellectual life and the survival of Banach, Poland's most influential mathematician. Was Banach spared because he had too much fame? Or were the occupying forces so mathematically illiterate that they had never heard of Banach? The author does not even speculate about these questions that beg to be answered.

As another example of a tantalizing tidbit from the book that needs more explanation, consider the following account (p. 51) of Banach's support for the

mathematical logician Leon Chwistek:

> when at some point Chwistek applied for a position in logic in Lvov, Banach backed him unequivocally and helped him to obtain the post. The affair scandalized half of intellectual Poland since Chwistek, in addition to being a respected scholar, also had a well-deserved reputation as being a somewhat strange and very eccentric artist.

Banach himself was "somewhat strange" and "eccentric"; that description surely fits many mathematicians. So why would Banach's support for such a person have "scandalized half of intellectural Poland"? Readers will realize that something more must have been involved here, but the author provides no hints to help solve this mystery.

In 1928 Stefan Banach and his colleague Hugo Steinhaus founded *Studia Mathematica*, which quickly became the most important journal specializing in the then new field of functional analysis. Today's mathematics librarians, grappling with budget problems, will be amused to learn that the first volume of *Studia Mathematica* cost \$1.50 outside Poland.

When teaching the graduate course in functional analysis, I always use the Krein-Milman Theorem and its appearance in *Studia Mathematica* as an excuse to inject a bit of history into the classroom. The Krein-Milman Theorem states that in a locally convex topological vector space, every compact convex set is the closed convex hull of its extreme points. This result was published (in somewhat less generality than the version just stated) in the 1940 volume of *Studia Mathematica*, which also contained two papers written by Banach. That volume of the journal was printed on poor-quality paper, clearly due to wartime conditions. The most curious feature of the 1940 volume is that each article (they are all written in either English, French, or German) appears with an abstract in Russian. Obviously Lvov, where *Studia Mathematica* was published, lay in the Soviet zone of occupation at the time of publication. Two weeks after Germany had invaded Poland from the west in September 1939, the Soviet Union marched into Poland from the east. Poland was partitioned between Germany and the Soviet Union until the summer of 1941, when Germany attacked the Soviet Union and occupied all of Poland.

The 1940 volume of *Studia Mathematica* was the last one edited by Banach, who died at age 53 shortly after World War II ended in 1945. After an absence of eight years, *Studia Mathematica* resumed publication in 1948 in Wrocław. Poland's border had moved westward after World War II, so that Lvov was then in the Soviet Union (no doubt this accounts for the *Britannica's* claim that Banach was a "Soviet mathematician"). A few years ago Lvov again changed countries—it is now part of Ukraine. Today *Studia Mathematica*, still a fine journal specializing in functional analysis, is published in Warsaw. The cover of each issue still proudly bears the names of the founding editors, Banach and Steinhaus.

In 1932 Banach published his famous book *Théorie des Opérations Linéaires*, based on his Polish version published a year earlier. Remarkably, *Théorie des Opérations Linéaires* remains in print today more than six decades after its original publication, partly because of its historic value as the first monograph on functional analysis but also because of the clean, modern style with which Banach presents the fundamentals of the subject (as created in good part by him and his collaborators). While a graduate student, I read *Théorie des Opérations Linéaires* to study for my French exam. I remember the thrill of seeing functional analysis developed by a legendary hero of twentieth century mathematics and my delight in

his extraordinarily clear writing. I also remember my amusement that what we today call "Banach spaces" are called "spaces of type (B)" in Banach's book. From the book under review I learned that Banach had previously written several popular high school mathematics textbooks for use throughout Poland; perhaps writing for a high school audience had honed Banach's excellent expository skills.

The Life of Stefan Banach left me hungry for more information about this fascinating figure. However, the author has performed a valuable service by uncovering some previously unknown data about Banach and by interviewing many of the dwindling number of people who knew Banach. This sketchy biography is a good place to start for someone wanting to learn about Banach.

Department of Mathematics
Michigan State University
East Lansing, MI 48824
axler@math.msu.edu

---

101 Careers in Mathematics. Edited by Andrew Sterrett. Mathematical Association of America, 1996, 250, $20.

Reviewed by J. Kevin Colligan

101 Careers in Mathematics is going to be ammunition for both sides of a lot of discussions. Andy Sterrett did a marvelous job of collecting an impressive variety of stories from people with degrees in mathematics and related disciplines. These people are now working in a stunning range of occupations. They include an astronaut, several college professors, marketing consultants, teachers at the elementary and secondary levels, mathematicians in government service (I have to put in a plug for my own "brand" of mathematician!), engineers, computer scientists, actuaries, a health professional, a Deputy Assistant Secretary of Defense, and even the Director of Inventory Control for L. L. Bean. Mathematics majors go everywhere and do everything.

On the whole, at least based on the 101-person sample in this book, they're fairly successful, too. And every one of them attributes his or her success and happiness to their experience in their undergraduate partial differential equations course!

If you believe that last statement, you need to get outside for a stronger dose of reality. They are in fact fairly successful and happy, but it wasn't the PDE course that did it, at least not exclusively. Some recurring themes in their comments stood out for me; I'll point them out to get you thinking about these items, and perhaps to start a discussion. Here's what I heard.

(0) There seemed to be a reasonable balance across gender and degree level.
(1) A lot of these people did, and continue to do, a lot of computer-related work.
(2) Many say the "mathematical way of thinking" or the crystallization of their thinking skills was the best gift they received from their education.
(3) Some—but not as many as I would have expected or hoped—said communication skills and people skills count.

(4) Several referred to the current need to perform in a teaming environment, especially cross-disciplinary ones.
(5) There was one reference to an expansion of individual and collective abilities: I have a few thoughts on this that I'll put off until the end.

Let's take these in turn.

(0) Gender and degree balance. I didn't go through the book building a 2-by-3 contingency table to compare degree level (roughly 3) against gender (roughly 2). There was a time when I might have done this, but my pile of reading material is too tall right now; if you decide to do it, let me know what comes out. It's balanced by gender, and I presume I'd see about the breakdown by degree level that recent statistics indicate: about half of the bachelor's degrees go to women, maybe 40% of the master's degrees, and about 25% of the doctorates. The fact that those graduate percentages are increasing so slowly continues to bother me. I've heard all the reasoned explanations, rationalizations, excuses, and finger-pointing. We have to put all that aside and solve the problem. No, I don't know the cure-all here, either, but some places are being successful: let's clone what works in those programs.

This book showcases a lot of successful women; so does *She Does Math!*, another excellent book on this topic, also published by the Association. Copies of these books should be in every post-secondary institution in the country. Maybe every high school, too. I can now hear you now telling me to get outside for a stronger dose of reality. Well, it's important to dream, isn't it? Let's use all the talent we have in this country.

(1) Computer-related work. Computers fly our planes, control our cars, run (or at least facilitate) our economy, run (no qualifier needed) our infrastructure, amuse us, teach us, and drive us nuts. Computers and knowing how to use them are integral parts of most if not all of the careers described in this book. We've made tremendous strides in using and applying this technology to all our jobs. We're still struggling to integrate them, to find the correct niche for technology.

At one end we have questions such as, should all knowledge be marketable? The results of the human genome project? Phone number databases? Purchase profile databases? Where does one draw the line? Our technological development has clearly outstripped our moral development. To make proper and prudent use of the technology we already have, we need to refine our understanding of the individual and collective ethical framework for our actions.

At the other—or at least nearer—end, how should we bring technology (primarily but not exclusively computers) into mathematics instruction? Students are going to use (will *have* to use!) these tools when they get a job. How are we fulfilling our ethical and professional duty to prepare them for their personal and professional lives? I've heard the argument that one loses proficiency by relying on a "computer crutch." In the October, 1996 issue of NCTM's *Mathematics Teacher*, Ken Ross said:

> Since it is easier to measure and spot deficiencies in skills than in understanding, this decline can easily be overemphasized. This problem is serious, however, especially since our future scientists, engineers, and mathematicians must obtain *both* substantial understanding and substantial skills. The reform movements need to address this issue.

Whether one is skilled at or challenged by mathematics, I'd much prefer having someone come away with an appreciation for how to approach a problem, and what resources can and ought be brought to bear to solve it. Frederick L. Frostic, a Deputy Assistant Secretary of Defense, said, "Those who know how things work always seem to work things best." Teach the skills; teach the technology: they'll need 'em both. The people in this book do.

(2) The mathematical way of thinking. Most of the *101 Careers* people said that their most valuable and utile acquisition as a mathematics major was, how to think. Sometimes we make mathematics look like a duck. We show the finished product, the calm and beauty above the waterline, without showing all the paddling underneath that gets us where we're headed. If we want to develop that mathematical way of thinking—what I believe Wade Ellis calls the HOTS (Higher Order Thinking Skills)—we'd better show some of the scaffolding along with the finished product. I have been told that George Mackiw, a friend of mine from Loyola College in Maryland, occasionally stops in class after completing a proof and says, "Now I want you to step back for a moment and appreciate the beauty of that proof." I like that. It encourages one to look at the edifice at different granularities. That's a core competency for a mathematician.

This way of thinking, the ability to both abstract and refine as needed, is one of the most valued gifts that instructors can give to all mathematics students. Teach them to analyze, to generalize, to look for patterns, but most of all, teach them to think. Ron Bosquet, R & D Product Manager for Hewlett-Packard, said, "My mathematics training ... taught me how to learn" and he is only one of many who say essentially the same thing.

(3) Communication and people skills. Some people in the book singled this out as being important. *Some* did, but not as many as I expected. Well, at least no one said it was unimportant. I would have put this first in my list: it's that important to me and to the environment I live in, and I speculate that it's implicitly part of everyone's environment.

Some of the 101 said that these skills enable them to communicate between technical and non-technical constituencies, and I say, hooray for them! I've seen and worked with people at both ends of this spectrum. The ones that succeed are those who can communicate well orally and in writing (at least one person said *graphically*, too), who can explain technical reasoning to a non-technical audience, who understand that priorities—and hence decisions—can be determined by views and feelings and not always by facts. These individuals are the ones whose voices matter, the ones that can be relied on to cut to the quick of complicated issues. They're the ones who make a difference. It's more fun to make a difference.

So how does this bear on what happens in the classroom? Well, let's think not only about how we personally give stellar examples of crystal-clear communication and interpersonal skills, but also about how we encourage this behavior in students. Are there group activities? Oral presentations? Writing assignments? Do we encourage and reward clarity and grammar? How about asking for an "executive summary" of a proof or argument?

Of course, all of these skills are interconnected. As professional mathematicians, we owe it to our colleagues, students, customers, or whomever to build and use these skills ourselves, and aid and abet their development in others.

(4) Teaming. This is another apple pie and motherhood issue. In my opinion, the days of Herculean solitary efforts are gone. (Okay, I have to think about how to finesse Andrew Wiles's work into this mold.) Has anyone done a count of the

average number of authors per paper submitted to journals lately? Has anyone studied this over time? Problems—at least most of the ones that the *101 Careers* people talked about—are multidisciplinary.

I like working on teams. I generally learn more mathematics, people thank me for providing the expertise I bring to the team, and I often get introduced to a whole new class of problems that I didn't even know about. Done right, teams create synergy that invigorates all team members.

It's not an innate skill that everyone's born with, though. The "parallel play" that is often seen in very young children sometimes translates into "parallel work" in adults, even in a supposedly team-oriented environment. It seems to be something we can be taught to do, however: there are quite a few entrepreneurs out there offering $800-a-pop courses that claim to develop teaming skills. Mathematics courses should develop teaming skills as well. Investigate, question, experiment, even guess: in a team that has developed mutual trust, these activities can accelerate learning and show different perspectives on both solutions and stumbling blocks. There have been many times when I thought I understood something, only to have someone else's question make me realize there was more going on than I originally saw. More than once, such experiences opened my eyes to structure I didn't even expect.

Teaming is useful: we can help teach it.

(5) Expansion of collective abilities. Here's a quotation from Sherra Kerns, Chair of the Electrical Engineering and Computer Science Department at Vanderbilt University: "Computerization [is] a facilitation and expansion of our individual and collective capabilities, a revolution at the edge of global opportunities affecting each touched life." I say, so is mathematics. [Soapbox alert!] It encourages us to develop global and local perspectives. It brings us the satisfaction of solving problems, of seeing patterns, often of simply fixing things, and it brings this to us both as individuals and in groups. In the words of Samson Cheung, a research scientist at NASA Ames, "Mathematics can provide one with the tools to use one's imagination."

I'm going to raise a moral issue here again. The unreasonable applicability of mathematics to the real world has an ethical dimension. Issues of access to mathematical training are not moot: mathematical knowledge empowers people. Harlan Mills, an industrial consultant, said "I have been continually surprised at the level of mathematics education and maturity in successful colleagues in apparently nonmathematical positions or activities." Likewise, mathematical ignorance can disenfranchise them. Like it or not, we are gatekeepers. That brings with it obligations to discern carefully and thoughtfully our proper role, and act on it.

All in all, I liked this book: it made me think about the breadth of my profession, and that forced me to think about its depth as well. The appendices, reprinted from *Math Horizons*, the MAA's student magazine (which you should also order), give great suggestions for job applicants and interviewers alike. Get a copy of *101 Careers in Mathematics* to see what comes out the far end of the mathematics education pipeline. It's thought-provoking.

J. Kevin Colligan
National Security Agency
Fort George G. Meade, MD 20755-6709
jkev@romulus.ncsc.mil

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**

with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1–4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, P, L. *Insights of Genius: Imagery and Creativity in Science and Art.* Arthur I. Miller. Springer-Verlag, 1996, xxii + 482 pp, $27. [ISBN 0-387-94671-3] "To see is to understand." A thorough explanation of the role of visual images in prompting imagination and creativity, using 20th-century physics and art as case studies. Includes a fascinating discussion of the profile of creativity of Henri Poincaré prepared by his contemporary, the psychologist Edouard Toulouse. LAS

Reference, P. *Handbook of Brownian Motion—Facts and Formulae.* Andrei N. Borodin, Paavo Salminen. Prob. & Its Applic. Birkhäuser Boston, 1996, xiv + 462 pp, $129. [ISBN 0-8176-5463-1] First part summarizes theory of linear diffusions. Second part is table of distributions of functionals of Brownian motion and related processes.

Reference, P, L*. *A Compendium on Nonlinear Ordinary Differential Equations.* P.L. Sachdev. Wiley, 1997, xi + 918 pp, $125. [ISBN 0-471-53134-0] Information about closed form solutions, asymptotics, stability, existence, and numerical results.

Reference, P. *Learning LaTeX.* David F. Griffiths, Desmond J. Higham. SIAM, 1997, x + 84 pp, $15.50 (P). [ISBN 0-89871-383-8] A clear, simple, up-to-date, sometimes amusing introduction to LaTeX—brief yet surprisingly comprehensive. Covers basics, graphics, bibliographies, indexes, slides, electronic resources, differences between "old" and "new" LaTeX, etc. Rich with brief but pertinent examples. PZ

Recreational Mathematics, P. *Intriguing Mathematical Problems.* Oswald Jacoby, William H. Benson. Dover, 1996, x + 191 pp, $6.95 (P). [ISBN 0-486-29261-4] Republication of *Mathematics for Pleasure* (McGraw-Hill, 1962).

Recreational Mathematics. *Math and Logic Puzzles for PC Enthusiasts.* J.J. Clessa. Dover, 1996, x + 131 pp, $5.95 (P). [ISBN 0-486-29192-8] A slightly altered republication of *Micropuzzles* (Pan Books, 1983). 135 puzzles; some intended to be done using a computer.

Education, P, L. *Characterizing Pedagogical Flow: An Investigation of Mathematics and Science Teaching in Six Countries.* William H. Schmidt, *et al.* Kluwer Academic, 1996, xiv + 229 pp, $110. [ISBN 0-7923-4272-0] Results from a pre-TIMSS investigation of typical classroom practice of mathematics teachers in grades 4 and 8. Chief conclusions: In France and Spain teachers organize and present formal, complex subject matter; in Norway and Switzerland classes are characterized by student exploration; in Japan the emphasis is on multiple approaches to carefully selected examples; and in the United States teachers present information and direct student activities. LAS

Education, P. *Bold Ventures, Volume 3: Case Studies of U.S. Innovations in Mathematics Education.* Eds: Senta A. Raizen, Edward D. Britton. Kluwer Academic, 1996, xiii + 376 pp, $140. [ISBN 0-7923-4233-X] Extensive analysis of three reform efforts in mathematics education: the development of standards by NCTM, the innovative modelling-based precalculus course developed by the North Car-

olina School of Science and Mathematics, and the Urban Mathematics Collaboratives established by the Ford Foundation. Based on interviews with participants and observers, as well as visits to programs and classrooms. Intended as case studies in effecting change. LAS

History, S(16–18), P. *Mathematical Encounters of the Second Kind.* Philip J. Davis. Birkhäuser Boston, 1997, viii + 304 pp, $24.95. [ISBN 0-8176-3939-X] Encounters of the "second kind" are with the people of mathematics (the "first kind" being reserved for mathematics itself). Four meandering reminiscences, with some fiction mixed in, about emperor Napoleon's gift to mathematics, mathematician Stefan Bergman, and biochemist Lord Victor Rothschild. Written in Davis's engaging personal style. LAS

Foundations, T(18: 1), P. *Vicious Circles: On the Mathematics of Non-Wellfounded Phenomena.* Jon Barwise, Lawrence Moss. CSLI Lect. Notes, No. 60. Center for the Study of Language & Information (Leland Stanford Junior Univ, Stanford, CA 94305) & Cambridge Univ Pr, 1996, x + 390 pp, $24.95 (P); $49.95. [ISBN 1-57586-008-2; 1-57586-009-0] An introduction to hypersets, a concept that transcends sets. Explains recent results from mathematics, computer science, and philosophy. Includes applications to game theory, graph theory, semantical paradoxes, automata, and languages. DB

Foundations, P. *Kreiseliana: About and Around Georg Kreisel.* Ed: Piergiorgio Odifreddi. AK Peters, 1996, xiii + 495 pp, $60. [ISBN 1-56881-061-X] Essays for Georg Kreisel's 70th birthday: reminiscences, explanations of his work in logic and philosophy, technical papers in logic. DB

Combinatorics, P. *Embeddability in Graphs.* Liu Yanpei. Math. & Its Applic. (China Ser.). Kluwer Academic, 1995, xvi + 398 pp, $198. [ISBN 0-7923-3648-8]

Number Theory, P. *Introduction to Cyclotomic Fields, Second Edition.* Lawrence C. Washington. Grad. Texts in Math., V. 83. Springer-Verlag, 1997, xiv + 487 pp, $59. [ISBN 0-387-94762-0] New edition of this classic on p-adic L-functions includes recent work of Thaine, Kolyvagin, and Rubin, as well as application of Jacobi sums to primality testing. (*First Edition*, TR, February 1983.) DB

Group Theory, T(18: 1), P. *3–Transposition Groups.* Michael Aschbacher. Tracts in Math., V. 124. Cambridge Univ Pr, 1997, vii + 260 pp, $49.95. [ISBN 0-521-57196-0] Part I contains the first published proof of Fischer's classification of almost simple groups generated by 3–transpositions. Parts II and III establish structural results on Fischer groups and provide a foundation for the theory of sporadic groups. TH

Group Theory, P. *Symplectic Fibrations and Multiplicity Diagrams.* Victor Guillemin, Eugene Lerman, Shlomo Sternberg. Cambridge Univ Pr, 1996, xiv + 222 pp, $49.95. [ISBN 0-521-44323-7]

Group Theory, P. *Symmetric Inverse Semigroups.* Stephen Lipscomb. Math. Surveys & Mono., V. 46. AMS, 1996, xviii + 166 pp, $49. [ISBN 0-8218-0627-0]

Algebra, P. *Abelian Groups and Modules.* Eds: David M. Arnold, Kulumani M. Rangaswamy. Lect. Notes in Pure & Appl. Math., V. 182. Marcel Dekker, 1996, xii + 411 pp, $165 (P). [ISBN 0-8247-9789-2] Proceedings of a 1995 conference held in Colorado Springs, Colorado.

Algebra, P. *Semigroup Theory and its Applications.* Eds: Karl H. Hofmann, Michael W. Mislove. London Math. Soc. Lect. Note Ser., V. 231. Cambridge Univ Pr, 1996, ix + 165 pp, $29.95 (P). [ISBN 0-521-57669-5] Invited survey papers from a 1994 conference at Tulane University.

Algebra, P. *Torsion Theories Over Commutative Rings.* Willy Brandal, Erol Barbut. BCS Associates, 1996, ix + 112 pp, $28 (P). [ISBN 0-914351-06-0]

Algebra, P. *Semigroups and Their Subsemigroup Lattices.* L.N. Shevrin, A.J. Ovsyannikov. Math. & Its Applic., V. 379. Kluwer Academic, 1996, xi + 378 pp, $173. [ISBN 0-7923-4221-6]

Algebra, P. *Infinite-Dimensional Lie Groups.* Hideki Omori. Transl: Hideki Omori. Transl. of Math. Mono., V. 158. AMS, 1997, xii + 415 pp, $99. [ISBN 0-8218-4575-6]

Calculus, S*(13), C. *Animating Calculus: Mathematica Notebooks for the Laboratory.* Ed Packel, Stan Wagon. Springer-Verlag, 1997, xv + 292 pp, $34.95 (P), with disk. [ISBN 0-387-94748-5] Republication of 1994 WH Freeman edition (TR, June–July 1995).

Calculus, T(13–14: 3). *Calculus, Seventh Edition.* Dale Varberg, Edwin J. Purcell. Prentice Hall, 1997, xv + 975 pp. [ISBN 0-13-518911-X] Modest revision. Includes more technology-based problems and projects.

Calculus, S*(14). *Mathematica Projects for Vector Calculus.* Michael M. Neumann, T. Len Miller. Kendall/Hunt, 1996, viii + 167 pp,

$19.45 (P). [ISBN 0-7872-2858-3] Projects on curves in the plane and in space, on functions of several variables, and on parameterized surfaces. First three chapters introduce *Mathematica* and present detailed solutions to two sample projects. Final chapter provides hints (or solutions) for about half of the projects. AO

Real Analysis, S*(14–16), P, L*. *A Primer of Real Functions, Fourth Edition.* Ralph P. Boas. Revised: Harold P. Boas. Carus Math. Mono., No. 13. MAA, 1996, xiv + 305 pp, $32.95. [ISBN 0-88385-029-X] This edition adds a chapter on integration. (*Third Edition*, TR, December 1982; Extended Review, March 1984.)

Complex Analysis, P. *Theory of Functions, Parts I and II.* Konrad Knopp. Transl: Frederick Bagemihl. Dover, 1996, vii + 150 pp, $8.95 (P). [ISBN 0-486-69219-1] Republication in one volume of Parts I and II.

Partial Differential Equations, P. *Theory and Applications of Partial Functional Differential Equations.* Jianhong Wu. Appl. Math. Sci., V. 119. Springer-Verlag, 1996, x + 429 pp, $59.95. [ISBN 0-387-94771-X]

Partial Differential Equations, P. *Elliptic Boundary Value Problems in the Spaces of Distributions.* Yakov Roitberg. Math. & Its Applic., V. 384. Kluwer Academic, 1996, xi + 415 pp, $219. [ISBN 0-7923-4303-4]

Partial Differential Equations, P. *Numerical Approximation of Hyperbolic Systems of Conservation Laws.* Edwige Godlewski, Pierre-Arnaud Raviart. Appl. Math. Sci., V. 118. Springer-Verlag, 1996, viii + 509 pp, $59.95. [ISBN 0-387-94529-6]

Dynamical Systems, T(16–18), P. *Discrete Hamiltonian Systems: Difference Equations, Continued Fractions, and Riccati Equations.* Calvin D. Ahlbrandt, Allan C. Peterson. Texts in Math. Sci., V. 16. Kluwer Academic, 1996, xiv + 374 pp, $175. [ISBN 0-7923-4277-1] Thorough introduction includes discussion of the relationship of continued fractions to solutions of Riccati equations. Other topics include linear symplectic systems, discrete variational theory, symmetric three-term recurrences, and Bohner theory. MPR

Dynamical Systems, P. *Dynamic Systems on Measure Chains.* V. Lakshmikantham, S. Sivasundaram, B. Kaymakcalan. Math. & Its Applic., V. 370. Kluwer Academic, 1996, x + 285 pp, $144. [ISBN 0-7923-4116-3]

Numerical Analysis, P, L. *Numerical Recipes in Fortran 90: The Art of Parallel Scientific Computing, Second Edition.* William H. Press, *et al.* Cambridge Univ Pr, 1996, xx + 551 pp, $44.95. [ISBN 0-521-57439-0] Program listings, with hints on parallelization, for the Fortran 90 version of the Numerical Recipes library. Also includes an introduction to Fortran 90 and parallel programming. AO

Operator Theory, P. *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations.* R.E. Showalter. Math. Surv. & Mono., V. 49. AMS, 1997, xiii + 278 pp, $75. [ISBN 0-8218-0500-2]

Operator Theory, P. *The Asymptotic Distribution of Eigenvalues of Partial Differential Operators.* Yu. Safarov, D. Vassiliev. Transl. of Math. Mono., V. 155. AMS, 1997, xiii + 354 pp, $119. [ISBN 0-8218-4577-2]

Functional Analysis, P. *Nonlinear Integral Equations in Abstract Spaces.* Dajun Guo, V. Lakshmikantham, Xinzhi Liu. Math. & Its Applic., V. 373. Kluwer Academic, 1996, viii + 341 pp, $169. [ISBN 0-7923-4144-9]

Functional Analysis, T(18), S, P, L. *Numerical Range: The Field of Values of Linear Operators and Matrices.* Karl E. Gustafson, Duggirala K.M. Rao. Universitext. Springer-Verlag, 1997, xiv + 189 pp, $34.95 (P). [ISBN 0-387-94835-X] Very readable, well-written introduction. Each chapter concludes with interesting notes and references. Highly recommended. Undergraduates attracted by linear algebra will enjoy this book. KS

Functional Analysis, T(16–17), S, L. *An Introduction to Functional Analysis in Computational Mathematics.* V.I. Lebedev. Birkhäuser Boston, 1997, x + 255 pp, $59.95. [ISBN 0-8176-3888-1] Basic functional analysis, approximation theory, linear operator theory, iteration methods for solutions to operator equations. English translation of 1987 Russian text has missing articles, awkward sentence constructions, spelling errors. KS

Functional Analysis, P. *Distortion Theorems in Relation to Linear Integral Operators.* Yûsaku Komatu. Math. & Its Applic., V. 385. Kluwer Academic, 1996, viii + 305 pp, $147. [ISBN 0-7923-4304-2]

Analysis, P. *Best Approximation by Linear Superpositions (Approximate Nomography).* S. Ya. Khavinson. Transl. of Math. Mono., V. 159. AMS, 1997, vii + 175 pp, $69. [ISBN 0-8218-0422-7]

Analysis, T(17), P. *An Introduction to the Mathematical Theory of Inverse Problems.* Andreas Kirsch. Appl. Math. Sci., V. 120. Springer-Verlag, 1996, x + 282 pp, $59.95. [ISBN 0-387-94530-X] Basic properties of

regularization methods for linear ill-posed problems and an introduction to nonlinear inverse problems (spectral theory and scattering theory). Assumes background in real analysis. AO

Algebraic Geometry, T(16–17: 1, 2), L.** *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra, Second Edition.* David Cox, John Little, Donal O'Shea. Undergrad. Texts in Math. Springer-Verlag, 1997, xiii + 536 pp, $45. [ISBN 0-387-94680-2] Very nice introduction to modern algebraic geometry. Nicely worked out examples, rich exercises, appendices on how to use Mathematica, Maple, Axiom, Reduce, etc. (*First Edition*, TR, August–September 1993; Extended Review, June–July 1994.) RM

Differential Geometry, P. *Global Analysis in Mathematical Physics: Geometric and Stochastic Methods.* Yuri Gliklikh. Transl: Viktor L. Ginzburg. Appl. Math. Sci., V. 122. Springer-Verlag, 1997, xv + 213 pp, $59.95. [ISBN 0-387-94867-8] Unified treatment of problems in differential geometry and classical mechanics, stochastic differential geometry and statistical and quantum mechanics, and infinite differential geometry of groups of diffeomorphisms and hydrodynamics, effected by the use of the Newton equation and its analogs. RM

Geometry, P. *Elliptic and Parabolic Methods in Geometry.* Eds: Ben Chow, *et al.* AK Peters, 1996, vi + 203 pp, $59. [ISBN 1-56881-064-4] 12 papers from a 1994 workshop at the Geometry Center at the University of Minnesota.

Algebraic Topology, P. *Rings, Modules, and Algebras in Stable Homotopy Theory.* A.D. Elmendorf, *et al.* Math. Surveys & Mono., V. 47. AMS, 1997, xi + 249 pp, $62. [ISBN 0-8218-0638-6]

Algebraic Topology, P. *Equivariant Homotopy and Cohomology Theory.* J.P. May, *et al.* CBMS Reg. Conf. Ser. in Math., No. 91. AMS, 1996, xiii + 366 pp, $49 (P). [ISBN 0-8218-0319-0] Expository lectures from a 1993 NSF–CBMS conference at the University of Alaska.

Operations Research, P. *Designing Innovations in Industrial Logistics Modelling.* Eds: Andrew Kusiak, Maurizio Bielli. Math. Modelling Ser. CRC Pr, 1997, 268 pp, $69.95. [ISBN 0-8493-8335-8] 12 papers from a 1993 workshop held in Capri, Italy.

Operations Research, P. *Polling Systems.* S.C. Borst. CWI Tract, V. 115. Stichting Mathematisch Centrum, 1996, v + 232 pp, Dfl. 50 (P). [ISBN 90-6196-467-9]

Optimization, P. *Qualitative Topics in Integer Linear Programming.* V.N. Shevchenko. Transl. of Math. Mono., V. 156. AMS, 1997, xiii + 146 pp, $69. [ISBN 0-8218-0535-5]

Optimization, S(18), P, L. *Mathematical Programs with Equilibrium Constraints.* Zhi-Quen Luo, Jong-Shi Pang, Daniel Ralph. Cambridge Univ Pr, 1996, xxiv + 401 pp, $54.95. [ISBN 0-521-57290-8] Topics: theory of exact penalization, MPEC constraint qualifications, 1st- and 2nd-order optimality conditions, iterative algorithms (penalty-based interior point, implicit programming, piecewise sequential quadratic programming). KB

Optimal Control, P. *Control of Systems with Aftereffect.* V.B. Kolmanovskiĭ, L.E. Shaĭkhet. Transl. of Math. Mono., V. 157. AMS, 1996, xi + 336 pp, $99. [ISBN 0-8218-0374-3]

Probability, S(17), P, L. *Probability Theory and Combinatorial Optimization.* J. Michael Steele. CBMS–NSF Reg. Conf. Ser. in Appl. Math., V. 69. SIAM, 1997, viii + 159 pp, $26.50 (P). [ISBN 0-89871-380-3] Cutting edge probability theory directly applicable to combinatorial optimization. Three fundamental probabilistic themes: exploitation of martingales, use of subadditivity, application of Talagrand's isoperimetric theory of concentration inequalities. KB

Stochastic Processes, P. *Ergodicity for Infinite Dimensional Systems.* G. Da Prato, J. Zabczyk. London Math. Soc. Lect. Note Ser., V. 229. Cambridge Univ Pr, 1996, xi + 339 pp, $44.95 (P). [ISBN 0-521-579007]

Stochastic Processes, P. *Matrix-Analytic Methods in Stochastic Models.* Eds: Srinivas R. Chakravarthy, Attahiru S. Alfa. Lect. Notes in Pure & Appl. Math., V. 183. Marcel Dekker, 1997, xi + 375 pp, $150 (P). [ISBN 0-8247-9766-3] Proceedings of a conference in Flint, Michigan.

Stochastic Processes, T?(16–17: 1), P. *Competitive Markov Decision Processes.* Jerzy Filar, Koos Vrieze. Springer-Verlag, 1997, xii + 393 pp, $69. [ISBN 0-387-94805-8] Unified treatment of stochastic games (processes with two or more coupled controllers) and Markov decision processes (viewed as special cases of single decision maker, non-competitive games). Discussed from the perspective of mathematical programming, restricted to discrete time, finite state, Markov decision models with complete information. RM

Mathematical Computing, P. *Mathematica 3.0: Standard Add-on Packages.* Ed: Emily Martin. Cambridge Univ Pr, 1996, viii + 516 pp,

$29.95 (P); $59.95. [ISBN 0-521-58585-6; 0-521-58586-4] Descriptions of the functions included in the standard add-on packages.

Computer Science, P. *Computer Facial Animation.* Frederic I. Parke. AK Peters, 1996, xv + 365 pp, $59.95. [ISBN 1-56881-014-8]

Computer Science, P, L. *Applications on Advanced Architecture Computers.* Ed: Greg Astfalk. SIAM, 1996, xvii + 359 pp, $35 (P). [ISBN 0-89871-368-4] Collection of articles originally published in *SIAM News* between March 1990 and June 1995. Many of the articles have been updated or revised.

Computer Science, P. *Problems of Reducing the Exhaustive Search.* Eds: V. Kreinovich, G. Mints. AMS Transl., Ser. 2, V. 178. AMS, 1997, x + 189 pp, $79. [ISBN 0-8218-0386-7] 12 papers on propositional satisfiability and related problems.

Applications (Communication Theory), P. *Signals & Systems, Second Edition.* Alan V. Oppenheim, Alan S. Willsky, S. Hamid Nawab. Signal Proc. Ser. Prentice Hall, 1997, xxx + 957 pp. [ISBN 0-13-814757-4] (*First Edition*, TR, June–July 1983.)

Applications (Economics), T(16: 1), P, L. *Financial Calculus: An Introduction to Derivative Pricing.* Martin Baxter, Andrew Rennie. Cambridge Univ Pr, 1996, ix + 233 pp, $39.95. [ISBN 0-521-55289-3] Introduces basic ideas in discrete-time setting, then generalizes to continuous time. Applications to actual financial instruments and the interest rate market. AO

Applications (Engineering), C, P. *Fourier-Related Transforms, Fast Algorithms and Applications.* Okan K. Ersoy. Prentice Hall, 1997, xix + 522 pp, with disk. [ISBN 0-13-624412-2]

Applications (Engineering), P. *Mathematical and Numerical Modelling in Electrical Engineering Theory and Applications.* Michal Křížek, Pekka Neittaanmäki. Math. Mod.: Theory & Applic., V. 1. Kluwer Academic, 1996, xiii + 300 pp, $149. [ISBN 0-7923-4249-6]

Applications (Physical Science), P. *Mathematics of Microstructure Evolution.* Eds: Long-Qing Chen, *et al.* SIAM, 1996, ix + 391 pp, $50 (P). [ISBN 0-89871-386-2] 31 papers from a 1995 symposium in Cleveland, Ohio.

Applications (Physics), T(18: 2), P. *Conformal Field Theory.* Philippe Di Francesco, Pierre Mathieu, David Sénéchal. Grad. Texts in Contemp. Physics. Springer-Verlag, 1997, xxi + 890 pp, $89. [ISBN 0-387-94785-X]

Applications (Physics), P. *Geometry and Physics.* Eds: Jørgen Ellegaard Andersen, *et al.* Lect. Notes in Pure & Appl. Math., V. 184. Marcel Dekker, 1997, xxii + 745 pp, $185 (P). [ISBN 0-8247-9791-4] Invited papers from a series of four workshops, a summer school, and a conference held in 1995 in Denmark.

Applications (Physics), S(15–17). *Mathematics for Physicists.* Philippe Dennery, André Krzywicki. Dover, 1995, xiii + 384 pp, $12.95 (P). [ISBN 0-486-69193-4] Slightly corrected republication of 1967 Harper & Row text (TR, August–September 1968).

Applications (Systems Theory), P. *Stability Theory.* Eds: R. Jeltsch, M. Mansour. Intern. Ser. of Num. Math., V. 121. Birkhäuser Boston, 1996, vii + 249 pp, $122.95. [ISBN 0-8176-5474-7] Papers from a 1995 conference in Ascona, Switzerland marking the centennial of Hurwitz' paper on the location of roots of a polynomial.

Applications, T(17–18), S, L. *Distributions in the Physical and Engineering Sciences, Volume 1: Distributional and Fractal Calculus, Integral Transforms and Wavelets.* Alexander I. Saichev, Wojbor A. Woyczyński. Appl. & Num. Harm. Anal. Birkhäuser Boston, 1997, xviii + 336 pp, $45. [ISBN 0-8176-3924-1] A modern course in 'Advanced Mathematics for Engineers and Scientists.' Unifying theme is distribution theory, enriched with topics such as wavelets, nonlinear phenomena, white noise theory. Background needed in elementary differential equations, linear algebra, Fourier series, complex variables. Knowledge of Mathematica, Maple, or MATLAB useful for student projects. Well-written; examples, exercises with solutions. KS

Applications, T*(15–18). *Applied Mathematics, Second Edition.* J. David Logan. Wiley, 1997, xiv +476 pp, $69.96. [ISBN 0-471-16513-1] Solid follow-up to a very good *First Edition* (TR, April 1988). Major changes include additions to perturbation methods, division of ODE's and PDE's into different chapters, and improvements to coverage of integral equations. Chapters on similarity and finite difference methods have been deleted. Overall, an outstanding choice of topics. A fine book. MPR

Reviewers

KB: Karla Ballman, Macalester; DB: David Bressoud, Macalester; TH: Tom Halverson, Macalester; RM: Richard Molnar, Macalester; AO: Arnold Ostebee, St. Olaf; MPR: Matthew P. Richey, St. Olaf; KS: Karen Saxe, Macalester; LAS: Lynn Arthur Steen, St. Olaf; PZ: Paul Zorn, St. Olaf.

THE AUTHORS

RAJENDRA BHATIA is Professor at the Indian Statistical Institute, New Delhi. His research interests are in analysis of matrices and operators. He is the author of a new graduate text, *Matrix Analysis*, Springer-Verlag, New York, 1996.

PETER ŠEMRL received all his degrees from the University of Ljubljana, and is now Associate Professor at the University of Maribor. His research interests are functional analysis, operator theory, linear algebra, and functional equations.

ISTVÁN NEMES graduated from the Kossuth Lajos University in Debrecen, Hungary, and entered the Ph.D. program at the Research Institute for Symbolic Computation (RISC) of the Johannes Kepler University, Linz, Austria. His main interests are in elementary number theory and computer algebra. Working as a computer programmer, he finds time for solving problems proposed in this MONTHLY.

MARKO PETKOVŠEK graduated in technical mathematics from the University of Ljubljana, Slovenia, and has been associated with the Mathematics Department in Ljubljana ever since. He received his Ph.D. in computer science at Carnegie Mellon University in 1991 under the direction of Dana S. Scott. His interests include computer algebra, computer arithmetic, combinatorics, botany, mycology, and hiking.

HERBERT WILF has been Editor of this MONTHLY, and remembers well how the "Problems and Solutions" tail often wagged the MONTHLY dog. So he takes great pleasure in helping to solve 27 of them (problems, not dogs) at once.

DORON ZEILBERGER was born on July 2, 1950, in Haifa, Israel. He was born again (as a combinatorialist) on March 15, 1979, in Atlanta, Georgia. He got hooked on combinatorics right after reading 'A probablistic proof of the hook length formula' by Curtis Greene, et al. (Albert Nijenhuis), et al. (Herb Wilf) (*Adv. Math.* 31 (1979), 104–109). He lived happily ever after.

LAWRENCE W. BAGGETT received his Ph.D. from the University of Washington in 1966. He is Professor of Mathematics at the University of Colorado, Boulder. He has published many articles in the fields of ergodic theory and harmonic analysis. When working, he plays with wavelets; when playing, he prefers saltwater waves.

HERBERT A. MEDINA was born in El Salvador in 1964. He received his B.S. in Mathematics/Computer Science from UCLA, and his Ph.D. (1992) in Mathematics from UC Berkeley under the direction of Henry Helson. He is Assistant Professor of Mathematics at Loyola Marymount University. His research interests center on functional analysis, harmonic analysis, and ergodic theory; recently, he has taken an interest in mathematical biology. He also works on projects that aim to increase the participation of historically underrepresented groups in mathematics and the sciences.

KATHY D. MERRILL received her Ph.D. from the University of Colorado in 1983. She is a member of the Department of Mathematics at The Colorado College. Her principal research interests are ergodic theory, number theory, and their many rich interactions. When not studying a cocycle, she can be found riding a bicycle through the Colorado Rockies.

RICHARD A. MOLLIN is Professor of Mathematics at the University of Calgary. He obtained his Ph.D. in 1975 from Queen's University in Kingston, Ontario, where he was born. He returned to Queen's University five years later as one of

Canada's first NSERC University Research Fellows. Although his graduate training was in algebra, and he worked in this area for several years, he converted to the unarguably most important and beautiful area of mathematics—*Number Theory*. He has numerous publications in algebraic and computational number theory. The most recent is his book *Quadratics*, and he is writing a text on *Fundamental Number Theory with Applications* to be published in 1997.

WŁADYSŁAW KULPA born in 1944, received his Ph.D. in 1973 from the University of Silesia in general topology working under Professor Jerzy Mioduszewski. In 1974–75, he spent a year in Moscow as a participant in the Yu. M. Smirnov seminar on topology at the Lomonosov University. He has been Professor at the University of Silesia in Katowice, Poland, since 1993.

MELANIE WAHLBERG studied Spanish and mathematics as an undergraduate at Western Michigan University before completing a Master's degree in pure mathematics at Michigan State University. She returned to Western to pursue a Ph.D. in the teaching of undergraduate mathematics. Her research interests include calculus reform, writing in the mathematics classroom, and students' concept of proof. She enhances the pleasure of working on her Ph.D. with listening to classical music, playing word games, and enjoying outdoor activities in the beautiful surroundings of Michigan.

MORTAZA BAYAT received an M.S. degree in mathematics from the Institute for Advanced Studies in Basic Sciences and is continuing studies towards a Ph.D. His interests in mathematics are in number theory, analysis, functional analysis, and differential equations.

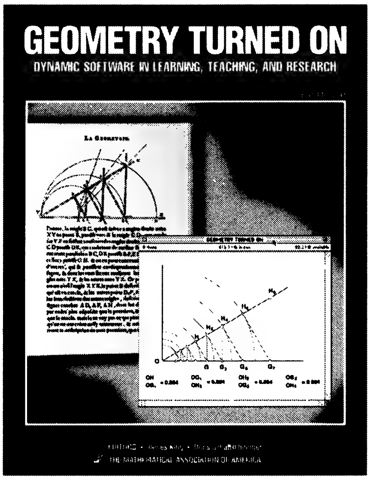
ZHANG BAO-LIN graduated from the Department of Mathematics of Peking University in 1964. He has worked in scientific computing since 1967. He is Professor, Laboratory of Computational Physics, Institute of Applied Physics and Computational Mathematics, Beijing, China. He has worked in parallel numerical mathematics and parallel computing for 10 years. He has taught graduate courses on approximation theory and parallel algorithms for many years inside and outside his institute.

JÓZSEF DÉNES graduated from Eötvös Loránd University in Budapest in 1954 and received his Ph.D. in mathematics from the same University in 1961. He is coathor of two books on Latin squares and has written more than 100 research papers on combinatorics, finite groups and semigroups, and cryptology. He was the Head of the Mathematical Department of the Institute for Co-ordination of Computing Techniques (SZKI) in Budapest. He retired in 1988 and is now a freelance industrial and scientific consultant.

SHELDON AXLER was an undergraduate at Princeton and a graduate student at Berkeley. He has served as an Associate Editor of the MONTHLY and as Editor-in-Chief of the *Mathematical Intelligencer*. He is Professor at Michigan State University. He has written two books: *Linear Algebra Done Right* and (jointly with two co-authors) *Harmonic Function Theory*. His nonmathematical activities include running and long-distance bicycling.

J. KEVIN COLLIGAN has been a mathematician at the National Security Agency since 1972. He is a Phi Beta Kappa, a Woodrow Wilson Fellow, a Billiken (St. Louis U.), and a Badger (Wisconsin). He enjoyed softball until his arm gave out, although it still has enough zip left to seriously hook a golf ball. He loves mathematics (especially number theory), algorithm analysis, and computer hardware; but he adores his New Orleans born and bred wife Suzanne and their six year old son Patrick, and the way both of them laugh.

THE MATHEMATICAL ASSOCIATION OF AMERICA



Geometry Turned On

Dynamic Software in Learning, Teaching and Research

James King and Doris Schattschneider, Editors

Learn how you can use interactive computer software in the study and teaching of geometry

Series: MAA Notes

Dynamic geometry is active, exploratory geometry carried out with interactive computer software. It has had a profound effect on classroom teaching wherever it has been introduced and has become an indispensable research tool for mathematicians and scientists. The papers in this volume give a good idea of the ways in which the software can be used, and some of the effects it can have. It is clear that the software raises various questions for teaching and research, and its continuing evolution raises questions on the design of the software itself.

With the use of interactive computer software, the focus in teaching shifts from students laboriously making constructions by hand to verify a stated fact in a text (for which there seems little reason to produce a proof) to a focus on students carrying out experiments, quickly producing many accurate sketches from which they conjecture properties that seem to be "always" true.

The latest dynamic geometry software takes advantage of current computer hardware with mouse interface for graphics and great speed. What is most exciting about the software is its dynamic nature. Thus, after a geometric configuration is drawn, any unconstrained parts of the configuration (arbitrary segments or points, for example), that are not dependent on any other objects are moveable — they can literally be grabbed with a cursor (using the mouse) and can be dragged or stretched — and as they move, all other objects in the configuration automatically self-adjust, preserving all dependent relationships and constraints.

Although this volume is printed in a conventional manner, and every paper has illustrations, most of the illustrations beg to be played with. We want you to be able to experience some of the explorations described by our authors. To make that possible, dynamic sketches that use Geometer's Sketchpad or Cabri II have been made available by several of the authors and are posted on a Web page maintained by the Mathematics Forum.

Catalog Code: NTE-41/JR97
275 pp., 1997, Paperbound
ISBN 0-88385-099-0
List: $38.95 MAA Member: $31.00

Phone in Your Order Now! ☎ 1-800-331-1622

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders**: $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ____________

Address ____________

City ____________ State _____ Zip ________

Phone ____________

Qty.	Catalog Code	Price	Amount
_______	NTE-41/JR97	_______	_______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling _______

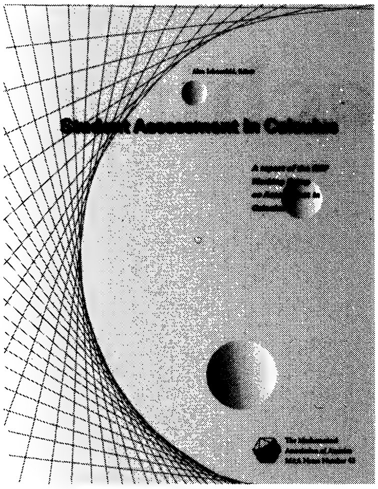
TOTAL _______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________ Expires ___/___

Signature ____________

THE MATHEMATICAL ASSOCIATION OF AMERICA



Student Assessment in Calculus

A Report of the NSF Working Group on Assessment in Calculus

ALAN SCHOENFELD, EDITOR

Series: MAA Notes

If you teach calculus, you should read this book. If you want to know what mathematics your students understand, or if you want to know how to find out what they understand, this book contains essential information for you.

It doesn't matter whether you teach a reform or traditional course, whether you have large or small sections, or whether you use lectures or laboratories. The bottom line is the same: When all is said and done, what counts is what our students understand. And that's what *Student Assessment in Calculus* is about.

Over the last ten years calculus instruction has changed in numerous ways. Whether they were trying on new ideas or following the more traditional routes towards conceptual understanding, both individual faculty and departments needed to know if their instruction was effective. To help deal with that issue, the National Science Foundation brought together a Working Group of experts in students' mathematical thinking, in assessment, and in calculus reform. The goals of their work were to:

- develop a framework to tailor calculus instruction to the students' needs;
- establish an agenda for further research on student understanding;
- describe how to make use of a range of techniques to test what students know, such as multiple-choice tests or short essay questions, student portfolios and "clinical" interviews;
- summarize major goals of the reform movement and describe the challenges faced by those who are taking a closer look at how students learn;
- illustrate the ways in which calculus projects attempt (via exams, papers, projects, etc.) to find out what their students have learned.

This book is the result of those efforts. If you teach calculus, if you want to see examples of useful assessment techniques, or if you are interested in issues of how to measure student learning in mathematics, then there is a lot for you here.

Catalog Code: NTE-43/JR97
122 pp., Paperbound, 1997
ISBN 0-88385-152-0
List: $34.95 MAA Member: $29.00

Phone in Your Order Now! ☎ 1-800-331-1622

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders**: $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ______________________

Address ______________________

City ______________ State _____ Zip ________

Phone ______________________

QTY.	CATALOG CODE	PRICE	AMOUNT
_______	NTE-43/JR97	_______	_______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling _______

TOTAL _______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

THE AMERICAN MATHEMATICAL MONTHLY

Volume 104, Number 7 August-September 1997

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and Notes should be sent to the Editor:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send your email address and 3 copies of the complete manuscript (including all figures with captions and lettering), typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper.

Letters to the Editor on any topic are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

See the MONTHLY section of MAA Online for current information such as contents of issues, descriptive summaries of forthcoming articles, tips for authors, and preparation of manuscripts in TEX:

http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all problems/solutions material, typewritten on only one side of the paper.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Mr. JOE WATSON, Advertising Manager

Change of address, missing issues inquiries, and other subscription correspondence:
MAA Service Center
maahq@maa.org

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT.

Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

The Many Avatars of a Simple Algebra

S. C. Coutinho

1. INTRODUCTION. Some mathematical structures show up in many different contexts, under many different guises. This is the case with the Weyl algebra. Born in the cradle of quantum theory, in the 1920s, it has come up in the representation theory of enveloping algebras and has played a key rôle in the creation of $\mathscr{D}$-module theory. It has recently returned to the parental home, under the auspices of deformation theory.

In this paper we survey the incarnations of the Weyl algebra associated to several formalisms of quantum mechanics. Beginning with the moment of conception in the 1920s, we work our way through matrix mechanics, Schrödinger's equation and Dirac's formalism. After a brief interlude where rings of differential operators are introduced, we return to quantum theory to look at *quantisation by deformation* and its version of the Weyl algebra.

2. QUANTUM MECHANICS. The story begins in May 1925, when W. Heisenberg fell ill with a bout of hay fever so vicious that he decided to ask for a fortnight leave to recover. He chose the island of Heligoland as a place to escape to. He must have been in a dreadful state indeed, because the landlady of the inn where he stopped for breakfast assumed, from his looks, that he had been involved in a fight the night before, [**21**, p. 248 ff].

In Heligoland, between walks and baths, Heisenberg carried on the work he had started in Göttingen. He was trying to develop a quantum mechanics, and his fundamental intuition was that it should deal only with observable quantities. Starting from that, Heisenberg developed a mathematical formulation of the theory. However it was not clear at first whether the mathematical scheme would be consistent or not.

Heisenberg felt that the real test of his scheme would be to check that it satisfied the law of conservation of energy. It took him a whole night to verify that energy was indeed conserved. Elated, he climbed a rock jutting out into the sea and watched the sun rise.

Let us see how Heisenberg arrived at his scheme of quantum mechanics. Consider an electron moving in an atom. If the system were classical, then we would have a function $x(t)$ describing the position of the electron as a function of time. We would also have Newton's equation

$$\ddot{x} + f(x) = 0.$$

Heisenberg decided that this equation ought to be retained, but that it would be necessary to find a new interpretation for $x(t)$. But the motion of the electron is periodic. Once again, if the system were classical, one could expand $x(t)$ as a Fourier series. In this case, the coefficients of the series would represent the amplitudes. In the quantum case these coefficients should depend on a quantum number. Developing the mathematical scheme along these lines, Heisenberg was led 'almost necessarily' to a very weird looking formula for the multiplication of

amplitudes. In particular, as he explicitly stated in his original paper, these amplitudes do not commute; a fact that deeply troubled him.

At first Heisenberg hoped to remove the need for non-commutative amplitudes from his theory. Unable to 'improve' the paper, he decided to come out with it and handed it over to Max Born shortly before leaving for England, where he would speak at the Kapitsa Club in Cambridge.

3. MATRIX MECHANICS. Born did not look immediately at Heisenberg's manuscript. It was the end of term, he felt tired and 'afraid of hard thinking' [**22**, p. 8 ff]. However, when he read through it a few days later, he was fascinated. Born immediately began to work on Heisenberg's ideas. By simplifying Heisenberg's notation and re-writing the formulae for the multiplication of amplitudes he immediately realised that it was formally like the product of matrices. It is interesting to note that at the time matrices were not in the toolkit of every physicist. Luckily Born still remembered matrices from his student days in Breslau, twenty years back.

Soon Born began his own 'constructive work'. Denoting by $\mathbf{p}$ and $\mathbf{q}$ the momentum and position variables of Heisenberg's picture, Born realised that $\mathbf{pq}$ and $\mathbf{qp}$ were different because $\mathbf{p}$ and $\mathbf{q}$ were matrices. He also noted that Heisenberg's formulae gave only the diagonal entries of the commutator $[\mathbf{p}, \mathbf{q}] = \mathbf{pq} - \mathbf{qp}$, which had to be $i\hbar$. Here $\hbar$ denotes Planck's constant divided by 2π.

In Born's own words: 'repeating Heisenberg's calculation in matrix notation, I soon convinced myself that the only reasonable value of the non-diagonal elements should be zero' [**27**, p. 37]. Thus he arrived at the formula

$$\mathbf{pq} - \mathbf{qp} = i\hbar\mathbf{1}, \tag{3.1}$$

where $\mathbf{1}$ denotes the identity matrix. In his words, this formula was 'only a guess, and my attempts to prove it failed'.

A few days later, Born met Pauli, on the train between Göttingen and Hanover. Unable to resist his enthusiasm, he told Pauli about his matrices and his difficulties with the proof of (3.1). Instead of showing interest, as Born had expected, Pauli accused him of spoiling Heisenberg's idea with 'futile mathematics' [**27**, p. 37].

Having failed to engage Pauli's interest, Born turned to his former student P. Jordan. Working together, they developed Heisenberg's idea in the context of matrix calculus. This is the first time that (3.1) appears in print, with a 'proof' due to Jordan [**27**, p. 277].

The version of quantum mechanics that follows from the work of Heisenberg, Born, and Jordan is called *matrix mechanics*. In it the *momentum* and *position* are represented by matrices. Denoting these matrices by $\mathbf{p}$ and $\mathbf{q}$, respectively, the *equations of motion* for an electron moving in one dimension, under a potential, take the form

$$\begin{aligned} \partial\mathbf{q}/\partial t &= \partial H/\partial\mathbf{p} \\ \partial\mathbf{p}/\partial t &= -\partial H/\partial\mathbf{q} \end{aligned} \tag{3.2}$$

where H is a function of $\mathbf{p}$ and $\mathbf{q}$. These are Hamilton's equations of motion, to which we return in the next section. The point to note here is that the equations involve two kinds of differentiation: by a scalar (time) and by matrices ($\mathbf{p}$ and $\mathbf{q}$). The first poses no problem, but the same cannot be said of the second. We return to this question in §5.

4. HAMILTONIAN MECHANICS. Let us briefly review a few facts about hamiltonian mechanics that we will require. Consider a particle of mass 1 moving along a straight line. Let q and p denote the position and momentum of the system. Since we have a classical system, these are numbers: the coordinates of phase space. Suppose that the particle is subject to a force $F(q, t)$, which depends on position and time.

Since the system is one dimensional, F can be derived from a potential $V(q, t)$, given by

$$V(q,t) = -\int_{q_0}^{q} F(q,t)\,dq.$$

Hence the total energy of the system, which is the sum of the potential and kinetic energy of the particle is

$$H(q,p,t) = \frac{p^2}{2} + V(q,t).$$

This is called the *Hamiltonian* or *Hamiltonian function* of the system. By Newton's second law

$$\frac{\partial p}{\partial t} = F(q,t) = -\frac{\partial H}{\partial q}.$$

On the other hand, a direct calculation shows that

$$\frac{\partial q}{\partial t} = p = \frac{\partial H}{\partial p}.$$

The equations

$$\begin{aligned} \frac{\partial p}{\partial t} &= -\frac{\partial H}{\partial q} \\ \frac{\partial q}{\partial t} &= \frac{\partial H}{\partial p}, \end{aligned} \tag{4.1}$$

are called *Hamilton's equations* of motion.

We have thus obtained Hamilton's equations for a system that consists of a particle of unit mass moving on a straight line under a force $F(q, t)$. In general, a *Hamiltonian system* of one degree of freedom is a second order system whose motion is determined by equations of the form (4.1).

The quantities of classical mechanics are described in terms of infinitely differentiable complex-valued functions of p and q. For the sake of simplicity we shall restrict ourselves to polynomial functions. Thus we shall be concerned with the space $\mathbf{C}[p, q]$ of polynomials in two commuting variables, which we denote by S. The system of equations (4.1) can be written in a very compact form using an operation called the *Poisson bracket* which is defined, for $f, g \in S$, by

$$\{f,g\} = \frac{\partial f}{\partial p}\frac{\partial g}{\partial q} - \frac{\partial f}{\partial q}\frac{\partial g}{\partial p}.$$

This is clearly a polynomial in p and q. The vector space S is a Lie algebra with respect to the Poisson bracket; but this will not be needed here.

Returning to (4.1), an easy calculation shows that if we write x for the vector (p, q), the equations can be re-written in the form

$$\dot{x} = \{H, x\} \tag{4.2}$$

where the bracket with H is calculated coordinatewise. At this stage this may seem just a little trick. In fact, the Poisson bracket is the algebraic counterpart of the symplectic structure that gives phase space its peculiar geometry; see **[2]**. Moreover, this formalism guided Dirac in his formulation of quantum mechanics, as we shall now see. For more details about the Hamiltonian formalism see **[26]** or **[1]**.

5 DIRAC. In the meantime, in Cambridge, Heisenberg mentioned his ideas on matrix mechanics at the end of his talk to the Kapitsa Club. One of the physicists present at the lecture was R. H. Fowler. In September 1925 Heisenberg sent the proofs of his paper to Fowler, who promptly handed them to his student Paul Dirac. Dirac looked at the paper but 'at first could not make much of it'. Returning to it two weeks later, he realised that it 'provided a clue to the problem of quantum mechanics'. Unaware of the developments in Germany, Dirac proceeded to work out his own version of quantum mechanics.

Instead of interpreting the quantum variables as matrices, Dirac calculated with them formally. To use the Hamiltonian formalism he had to find an interpretation for the operation of differentiation with respect to a quantum variable, as we have already observed at the end of §3. Dirac's solution was to point out that the quantum analogue of differentiation by $\mathbf{q}$ say, is taking the commutator with $\mathbf{p}$. Thus, if $\mathbf{f}$ is a function of $\mathbf{p}$ and $\mathbf{q}$ in the quantum algebra, then $d\mathbf{f}/d\mathbf{q} = [\mathbf{p}, \mathbf{f}]$.

One of Dirac's great contributions was his identification of the classical analogue of the quantum commutator. According to Bohr's *correspondence principle* the results of quantum mechanics should converge to the analogous classical results when Planck's constant tends to zero. This weird 'constant tends to zero' really means that the numerical value of the constant should be small when it is expressed in the units of action characteristic of the class of systems under consideration.

Guided by this principle, Dirac discovered that the commutator divided by $i\hbar$ is the quantum analogue of the Poisson bracket of classical mechanics. The analogy allowed him to derive formula (3.1). Furthermore, defining H to be the Hamiltonian of the system, and assuming that 'the orders of the factors of the products occurring in quantum motion are unimportant' he wrote the fundamental quantum equation in the form

$$\dot{x} = [H, x],$$

in complete analogy with (4.1).

This analogy also helps to explain Dirac's formula for differentiation by $\mathbf{q}$. Indeed, if f is a polynomial in the (commutative) variables p and q, one immediately checks from the formula of the Poisson bracket that $\{p, f\} = \partial f/\partial q$. The corresponding quantum formula is obtained by replacing $\{\,,\}$ with $[\,,]$.

As his papers show, Dirac clearly understood that the quantum mechanical quantities defined a new sort of algebra, for which the multiplication was not commutative. He later called these quantities *q-numbers*—as opposed to *c-numbers*, which are the ordinary complex numbers. Dirac's papers can be found in **[27**, p. 307 and 417**]**.

6. QUANTUM ALGEBRA. Let us consider the algebraic background of the two interpretations of quantum mechanics that we have surveyed. Dirac assumes that he has 'quantities' that behave in a certain way. In other words, *symbols* that are subject to *relations*. In the one dimensional case two symbols $\mathbf{p}$ and $\mathbf{q}$ are required to represent momentum and position. They are related by $\mathbf{pq} - \mathbf{qp} = i\hbar \cdot \mathbf{1}$—where

1 denotes the identity of the quantum algebra $\mathscr{A}$. To avoid unnecessary complication we *normalize* this relation to the form $\mathbf{pq} - \mathbf{qp} = \mathbf{1}$.

Algebraically, Dirac's quantum algebra is constructed beginning with the *complex free algebra* $\mathscr{F}$ in two generators x and y. The elements of the free algebra are linear combinations (with complex coefficients) of *words* in x and y. The product of two words is obtained by juxtaposition. The quantum algebra $\mathscr{A}$ is the quotient algebra of $\mathscr{F}$ by the two-sided ideal generated by $xy - yx - 1$. Thus $\mathbf{p}$ and $\mathbf{q}$ are the images of x and y in this quotient.

Every element of $\mathscr{A}$ is a linear combination of words in $\mathbf{p}$ and $\mathbf{q}$—a property that $\mathscr{A}$ inherits from its parent free algebra. Now, from $[\mathbf{p}, \mathbf{q}] = \mathbf{1}$ one deduces that

$$[\mathbf{p}, \mathbf{q}^k] = k\mathbf{q}^{k-1} \quad \text{and} \quad [\mathbf{p}^k, \mathbf{q}] = k\mathbf{p}^{k-1}$$

This agrees with Dirac's observation that commutation is analogous to differentiation. These commutation relations allow us to write every word in $\mathbf{p}$ and $\mathbf{q}$ as a linear combination of monomials $\mathbf{q}^k\mathbf{p}^m$. Thus every element of $\mathscr{A}$ is a linear combination of monomials of this form.

Working a little harder, we can show that the monomials $\mathbf{q}^k\mathbf{p}^m$, with $k, m \geq 0$, form a basis of $\mathscr{A}$ as a complex vector space [**6**, Proposition 1.2.1]. This can be used to define the degree of an element of $\mathscr{A}$. First define the *degree* of a monomial $\mathbf{q}^k\mathbf{p}^m$ to be $k + m$. Now write $\mathbf{d} \in \mathscr{A}$ as a linear combination of monomials of this form: the maximum of the degree of these monomials is called the degree of $\mathbf{d}$ and is denoted by $\deg(\mathbf{d})$.

The degree of $\mathscr{A}$ behaves in many ways like the degree of polynomials. For $\mathbf{d}_1, \mathbf{d}_2 \in \mathscr{A}$,

(1) $\deg(\mathbf{d}_1 + \mathbf{d}_2) \leq \max\{\deg(\mathbf{d}_1), \deg(\mathbf{d}_2)\}$,
(2) $\deg(\mathbf{d}_1\mathbf{d}_2) = \deg(\mathbf{d}_1) + \deg(\mathbf{d}_2)$, and
(3) $\deg[\mathbf{d}_1, \mathbf{d}_2] \leq \deg(\mathbf{d}_1) + \deg(\mathbf{d}_2) - 2$.

The proof of (1) is immediate, but the proof of (2) uses (3) and is somewhat convoluted. An immediate consequence of (2) is that $\mathscr{A}$ is an integral domain: it does not have any zero divisors. See [**6**, Ch. 2, §1].

The degree can be used to prove several properties of $\mathscr{A}$. For example, in §2 of [**7**], Dirac characterizes all the derivations of $\mathscr{A}$. Recall that a *derivation* D of $\mathscr{A}$ is a $\mathbf{C}$-linear operator of $\mathscr{A}$ that satisfies $D(\mathbf{d}_1\mathbf{d}_2) = \mathbf{d}_1 D(\mathbf{d}_2) + D(\mathbf{d}_1)\mathbf{d}_2$, for every $\mathbf{d}_1, \mathbf{d}_2 \in \mathscr{A}$. With Dirac, we note that the order of $\mathbf{d}_1$ and $\mathbf{d}_2$ in the formula cannot be changed. An easy way to produce derivations of $\mathscr{A}$ is to use the commutator. Given $\mathbf{f} \in \mathscr{A}$, define $P(\mathbf{d}) = [\mathbf{d}, \mathbf{f}]$, for $\mathbf{d} \in \mathscr{A}$. As one checks easily, this is a derivation of $\mathscr{A}$. Derivations of this form are called *inner derivations* of $\mathscr{A}$.

Dirac showed that all the derivations of $\mathscr{A}$ are inner. Let us sketch the proof. Let D be a derivation of $\mathscr{A}$. Since commutation by $\mathbf{p}$ and $\mathbf{q}$ behaves like differentiation, we can find $\mathbf{f} \in \mathscr{A}$ such that $D(\mathbf{p}) = [\mathbf{f}, \mathbf{p}]$ and $D(\mathbf{q}) = [\mathbf{f}, \mathbf{q}]$. The actual calculation is reminiscent of the way one finds a potential function for a conservative polynomial vector field on the plane. Now, using induction on the degree of $\mathbf{d} \in \mathscr{A}$, one can check that $D(\mathbf{d}) = [\mathbf{f}, \mathbf{d}]$.

Another very important property of $\mathscr{A}$ is that it has no proper two-sided ideals, except zero. In other words, $\mathscr{A}$ is a *simple* algebra. However, it is not a division ring: $\mathbf{p}$ cannot have an inverse because, when multiplied by any element of $\mathscr{A}$ it gives rise to an element of degree at least 1. Actually, the only invertible elements of $\mathscr{A}$ are the constants.

The proof that $\mathscr{A}$ is simple goes as follows: suppose that J is a non-zero two-sided ideal of $\mathscr{A}$. Choose a non-zero element $\mathbf{d} \in \mathscr{A}$. Commuting with $\mathbf{p}$ is

formally equivalent to differentiation by $\mathbf{q}$. Hence commuting $\mathbf{d}$ with $\mathbf{p}$ enough times we obtain an element $\mathbf{d}' \in \mathscr{A}$ that does not involve $\mathbf{q}$. But J is a two-sided ideal of $\mathscr{A}$. Thus every time we commute an element of J with $\mathbf{p}$, we get an element of J. Hence $\mathbf{d}' \in J$. Now repeat the process with $\mathbf{d}'$, this time commuting it as many times as necessary with $\mathbf{q}$, until we arrive at a non-zero constant. Thus J contains a non-zero constant, and so $J = \mathscr{A}$; this is what we wanted to prove. For details see [**6**, Theorem 2.2.1].

We can also describe in a mathematical way the relation that Dirac found between the quantum commutator and the Poisson bracket. Quantum mechanics is represented by $\mathscr{A}$, and classical mechanics is represented by the complex algebra S of polynomial functions on the variables p and q, which stand for momentum and position. Thus S is a commutative algebra.

Let B_k be the set of elements of $\mathscr{A}$ of degree $\leq k$ and let $S(k)$ be the set of homogeneous polynomials of degree k in S. We define a map σ_k: $B_k \to S(k)$ as follows: If $\mathbf{d} \in B_k$ has degree k, ignore the monomials of degree $< k$, and replace $\mathbf{p}$ by p and $\mathbf{q}$ by q in the monomials of degree k. This gives a homogeneous polynomial of degree k, which we denote by $\sigma_k(\mathbf{d})$. For example, if $\mathbf{d} = \mathbf{q}^4\mathbf{p}^5 + 7\mathbf{q}^3\mathbf{p}^6 + 6\mathbf{p}^5 + 3\mathbf{pq}$, then $\mathbf{d}$ has degree 9 and $\sigma_9(\mathbf{d}) = q^4p^5 + 7q^3p^6$. This is called the *symbol map* of degree k of $\mathscr{A}$; it is a linear map of vector spaces, no more. Note that if $\mathbf{d} \in B_k$ has degree $< k$ then its symbol of degree k is zero. This construction is well-known from partial differential equation theory.

Now to the relation with the Poisson bracket. Let $\mathbf{d}_1$, $\mathbf{d}_2$ be elements of $\mathscr{A}$ of degrees k_1 and k_2 respectively. By (2), the commutator $[\mathbf{d}_1, \mathbf{d}_2]$ has degree at most $k_1 + k_2 - 2$. One can now check that

$$\sigma_{k_1+k_2-2}([\mathbf{d}_1, \mathbf{d}_2]) = \{\sigma_{k_1}(\mathbf{d}_1), \sigma_{k_2}(\mathbf{d}_2)\}.$$

This is one way to express the relation discovered by Dirac. We will come across another way, more in the spirit of the correspondence principle, in §11.

7. MATRIX REPRESENTATIONS. Let us now turn to the Heisenberg-Born-Jordan version of quantum mechanics. In it $\mathbf{p}$ and $\mathbf{q}$ are matrices. First of all notice that there cannot be two finite matrices whose commutator is $\mathbf{1}$. The easiest way to see this is to observe that the trace of a commutator is always zero, while the trace of the identity matrix is always non-zero. Therefore any such matrices must be infinite.

Thus we are led to a representation of $\mathscr{A}$ into the algebra $M_\infty(\mathbf{C})$ of infinite matrices with complex coefficients. In other words, we must construct a homomorphism of algebras of $\mathscr{A}$ into $M_\infty(\mathbf{C})$. This is easy, given that we have defined $\mathscr{A}$ as a quotient of a free algebra. It is enough to find two matrices P and Q in $M_\infty(\mathbf{C})$ such that $PQ - QP = \mathbf{1}$; for example

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 2 & 0 & \cdots \\ 0 & 0 & 0 & 3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad \text{and} \quad Q = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

First define a map $\theta: \mathscr{F} \to M_\infty(\mathbf{C})$ by $\theta(x) = P$ and $\theta(y) = Q$. Since $PQ - QP = \underline{1}$, it follows that $xy - yx - 1$ belongs to the kernel of θ. Thus θ induces a map $\bar{\theta}$: $\mathscr{A} \to M_\infty(\mathbf{C})$. But we have already seen that $\mathscr{A}$ is simple. In particular, the image of $\ker(\theta)$ in $\mathscr{A}$ must be zero. Hence $\bar{\theta}$ is injective. In other words, the subalgebra of $M_\infty(\mathbf{C})$ generated by P and Q is isomorphic to $\mathscr{A}$.

Thus we have two ways of describing the quantum algebra $\mathscr{A}$: as a quotient of a free algebra (Dirac's way) or as a subalgebra of a matrix algebra (the Heisenberg-Born-Jordan way). A third scheme for doing quantum mechanics leads into yet one more description, perhaps the most fruitful, in terms of differential operators.

8. WAVE MECHANICS. It has long been known that light shows phenomena that are better explained in terms of waves, and others that make better sense if it is thought of as a stream of small particles. Quantum theory reached a compromise, affirming a dual nature for light, both wave and particle. In 1924, Louis de Broglie, then a student working towards his doctorate in Paris, understood that the wave-particle dualism ought to be truly universal. If that were so, then a 'particle' such as an electron should also present the same dual nature of wave and particle. Langevin sent a copy of de Broglie's thesis to Einstein, who wrote in reply: 'he has lifted a corner of the great veil'.

De Broglie's work was the starting point of a third version of quantum mechanics, developed by the Austrian physicist Erwin Schrödinger in 1925. Schrödinger's starting point can be best summed up in the aphorism *where there is a wave, there must also be a wave equation.* Actually this was reportedly said by P. Debye at the end of a colloquium in Zurich, in which Schrödinger explained de Broglie's work to his department; see [**23**, Ch. 6]. Using de Broglie's formulae and a reasonable heuristic argument, Schrödinger arrived at a very neat partial differential equation. For an electron of mass m moving in one dimension under a potential V, the equation is

$$i\hbar\frac{\partial\psi}{\partial t} = \frac{-\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} + V\psi. \tag{8.1}$$

At first this equation was hailed as a return to the good old days, with some physicists hoping that it would drive out those strange matrices and non-commutative quantities. Moreover, it was deterministic. However, what did the function ψ represent? It takes complex values, for a start. Max Born once again came to the rescue, and proposed that the *wave function*, as ψ came to be called, did not represent any physical quantity whatsoever. Only the square of its modulus $|\psi(x,t)|^2$ had a physical interpretation. It represented the probability of finding an electron at x at the moment t. Despite much initial dispute, this became the accepted interpretation.

This also rescued the *uncertainty principle*, which affirms that one cannot measure at the same time and with arbitrary precision, the position and momentum of a particle. In its mathematical form, it is a consequence of the quantum commutation relation (3.1). Since ψ is a solution of a differential equation, it behaves deterministically. But ψ cannot be measured. What one can measure is $|\psi|^2$, a mere probability.

Let us spell out this scheme in more detail. The wave functions live in the space $\mathscr{L}^2$ of square integrable functions defined on the real line and taking complex values. This is a Hilbert space. In particular, it is endowed with an inner product; if $\psi_1, \psi_2 \in \mathscr{L}^2$ then

$$\langle\psi_1,\psi_2\rangle = \int_{\mathbf{R}}\psi_1\overline{\psi_2}\,dx.$$

A wave function ψ must be square integrable because $\langle\psi,\psi\rangle$ is equal to the probability of finding the particle somewhere in the real line, which must be 1. However not every element of $\mathscr{L}^2$ is a wave function: wave functions must be

differentiable, if they are going to satisfy Schrödinger's equation. The *observables* correspond to Hermitian operators on $\mathscr{L}^2$. Despite the name, these cannot be observed directly. The magnitudes that are observed correspond to the eigenvalues of these operators. Since the observables are Hermitian operators, the eigenvalues are real numbers, which is what one would expect of physical quantities.

In the end, it turned out that Schrödinger's wave mechanics is equivalent to the matrix formulation. Note that in both versions one has operators: matrices, in matrix mechanics; differential operators, in wave mechanics. But in matrix mechanics, the matrices themselves change with time. The fundamental equations (3.2) relate the time derivative of a matrix with the quantum equivalent of the Hamiltonian. In wave mechanics, it is the wave function that changes with time. The differentiable operators act on the space of functions in the usual way.

The connection between the two pictures comes through an operator $U(t)$ that takes the wave function at t_0 into the wave function at t, namely $\psi(x,t) = U(t)\psi(x,t_0)$. It can be deduced from physical considerations that $U(t)$ is a unitary operator. Let X_0 be an observable in wave mechanics. Mathematically we are talking about a Hermitian operator in $\mathscr{L}^2$. Write $X = X(t) = U(t)^t X_0 U(t)$; this is the 'matrix' that corresponds to X_0 in matrix mechanics. Differentiating this formula with respect to t and using Schrödinger's equation, we arrive at

$$\dot{X}(t) = [H, X(t)],$$

which is the fundamental quantum equation in Dirac's form. For more details see [**8**, Ch. V, §28].

We can recreate the quantum algebra $\mathscr{A}$ in the language of wave mechanics. This time we will be handling differential operators in $\mathscr{L}^2$. The operators we want to consider are $\partial/\partial x$ and multiplication by x. For the sake of simplifying the formulae, let us denote these operators by ∂ and x, respectively. If ψ is a wave function, then

$$[\partial, x](\psi) = \partial(x\psi) - x\partial(\psi) = \psi.$$

Since this holds for every ψ, we conclude that $[\partial, x] = 1$, the identity operator in $\mathscr{L}^2$. Thus, proceeding as in §7, we can show that $\mathscr{A}$ is isomorphic to the complex subalgebra of $\mathrm{End}_{\mathbf{C}}(\mathscr{L}^2)$ generated by the operators ∂ and x. In wave mechanics, these operators correspond to momentum and position, as was to be expected.

9. $\mathscr{D}$-MODULES. We can represent the algebra $\mathscr{A}$ more economically as an algebra of differential operators if we use polynomial functions. Let us start in a little more generality. Let R be a commutative algebra over $\mathbf{C}$. We define the *ring of differential operators* $\mathscr{D}(R)$ inductively as a subalgebra of $\mathrm{End}_{\mathbf{C}}(R)$. Since an element of R gives rise to a linear operator of R by multiplication, the inductive definition begins with $\mathscr{D}^0(R) = R$, the operators of order zero. The operators of order k are

$$\mathscr{D}^k(R) = \{d \in \mathrm{End}_{\mathbf{C}}(R) : [d, a] \in \mathscr{D}^{k-1}(R) \text{ for all } a \in R\}.$$

Let $\mathscr{D}(R)$ be the union of all $\mathscr{D}^k(R)$ for $k \geq 0$. This turns out to be a subalgebra of $\mathrm{End}_{\mathbf{C}}(R)$, though the proof is not quite obvious, see [**6**, Ch. 3, §1].

It is easy to calculate $\mathscr{D}^1(R)$ explicitly. It is generated, as an R-module, by 1 and the $\mathbf{C}$-derivations of R. In particular, if $R = \mathbf{C}[x]$, the polynomial ring in one variable, then $\mathscr{D}^1(R) = R + R\partial$, where ∂ denotes the operator differentiation by x. Thus the quantum algebra $\mathscr{A}$ is contained in $\mathscr{D}(R)$ as the algebra generated by x and ∂. Working a little harder, we can prove that $\mathscr{A} = \mathscr{D}(R)$; for details see [**6**,

Ch. 3, §2]. Hence the quantum algebra is the algebra of differential operators of the ring of polynomials in one variable.

The preceding definition of rings of differential operators appears in Grothendieck's *Eléments de geométrie algébrique* [**14**, proposition 16.8.8]. The notoriety of rings of differential operators nowadays is mainly due to $\mathscr{D}$-module theory. A $\mathscr{D}$-module is a finitely generated module over the algebra of differential operators of the coordinate ring of a smooth affine algebraic variety. To handle general varieties one must introduce sheaves [**3**, Ch. VI].

The importance of the theory lies in its numerous applications, which extend from mathematical physics to number theory. One of the most famous is to the representation theory of algebraic groups, where $\mathscr{D}$-modules were used to settle the Kazhdan-Lusztig conjecture in 1981. A very important $\mathscr{D}$-module theoretic theorem used in the solution of the conjecture is the *Riemann-Hilbert correspondence*. This is a result of the noblest parentage. Its genealogical tree includes Riemann's memoir on the hypergeometric function, Hilbert's 21st problem, and the work of Deligne on regular connections.

10. THE WEYL ALGEBRA. It did not take long for algebraists to notice that the quantum algebra $\mathscr{A}$ was an interesting object of study. In 1933, D. E. Littlewood wrote a paper [**18**] in which he proves most of the properties of $\mathscr{A}$ that we considered in §6. He also gives several examples of infinite matrices satisfying (3.1), among them the one of §7.

Littlewood's language is rather antiquated. But in 1937, K. A. Hirsch published a paper [**16**] in which he proves that a class of rings that includes $\mathscr{A}$ is a simple algebra. This is a thoroughly modern paper, written in the language of van der Warden's *Moderne Algebra*. His approach is essentially the one presented in §6. See also [**5**].

A great boost to the study of $\mathscr{A}$ came with the realization that it appears as a quotient of enveloping algebras of nilpotent Lie algebras by primitive ideals. This brought them into the fold of the representation theory of Lie algebras.

In fact $\mathscr{A}$ is the first member of a family of complex algebras. It corresponds to a quantum system with one degree of freedom. The equations for systems with n degrees of freedom were found by Heisenberg himself, as early as September 1925. They also give rise to complex algebras that are simple integral domains. J. Dixmier studied these algebras in a series of papers in the 1960s, and was the first to call them *Weyl algebras*, after a suggestion of I. Segal; see [**10**]. He also introduced the notation $A_n(\mathbf{C})$ for the algebra corresponding to a system of n degrees of freedom; see [**9**]. Both the name and notation have become standard.

The importance of the Weyl algebra has grown steadily in the last 30 years; see [**3**], [**6**], [**20**]. The work on non-commutative Noetherian rings that followed A. Goldie's famous theorems on quotient rings of Noetherian rings [**12**], [**13**] and the fact that the Weyl algebra is the simplest (but quite typical) ring of differential operators has only added to its importance.

11. DEFORMATIONS. It is time to return to quantum mechanics. The three schemes that we studied in §§6–8 give rise to the method known as *canonical quantisation.* First of all, by *quantisation* we mean the process of turning a classical system into its corresponding quantum system. This is not a well-defined process. In canonical quantisation one starts with the Hamiltonian H of the classical system and systematically replaces the classical variables position and momentum by the

operators x and d/dx of wave mechanics. One may now write the corresponding Schrödinger equation and solve it.

Several other methods of quantisation have been proposed: geometrical quantisation, asymptotic quantisation, deformation quantisation. It is the last of these that we want to study here. It leads us into another way of describing the Weyl algebra: as a deformation of a polynomial ring.

Let S be a commutative $\mathbf{C}$-algebra. Denote by $S[[t]]$ the space of power series in one variable with coefficients on S. Note that we are considering $S[[t]]$ as a *vector space* only, and not as a ring. This is because what we really want to do is to define a new multiplication in $S[[t]]$. To do that, we start with a family of bilinear maps $B_j\colon S \times S \to S$, for $j \geq 0$. If $a, b \in S$, then their *$\star$-product* in $S[[t]]$ is

$$a \star b = \sum_{j=0}^{\infty} B_j(a,b)t^j.$$

Extending this linearly to the whole of $S[[t]]$, we obtain a multiplication in this space. The multiplication is associative if the B_j satisfy

$$\sum_{i+j=k} B_i\big(a, B_j(b,c)\big) = \sum_{i+j=k} B_i\big(B_j(a,b), c\big)$$

for $k \geq 0$ and all $a, b, c \in S$. This is not very easy to check for a given family of bilinear maps. Doing it recursively, one is led to consider Hochschild homology, as shown by M. Gerstenhaber in **[11]**. We do not pursue this line here; our aims are more modest.

Two further assumptions are usually made. Since we want the $\star$-product to be a *deformation* of S, we must have $B_0(a,b) = ab$, the original product in S. If the identity of S is to be the identity of $S[[t]]$ with the product $\star$, then we must also have $B_j(a,b) = 0$ for $j > 0$ if either a or b is a scalar.

Let us return to quantum theory. We have seen that one of the key features of Dirac's approach to quantum mechanics was the relation between the classical Poisson bracket and the quantum commutator. He arrived at this relation using the *correspondence principle*, which states that a quantum system should tend to its classical analogue when Planck's constant tends to zero. This is also the starting point of the deformation theoretic approach to quantum mechanics.

In this approach we begin with the classical phase space. Since we are considering only a particle moving in a straight line, phase space is a two dimensional space. The classical dynamical variables are functions on phase space, and we are assuming that they are polynomial functions, to keep the going easy. The same is true in the deformation theoretic scheme. So far, so good. What we have to define anew is the multiplication of these observables. Furthermore, it must somehow depend on Planck's constant.

So let $S = \mathbf{C}[p, q]$ be the polynomial ring. We define a new product in $S[[\hbar]]$, the space of formal power series in $\hbar$, using the deformation theoretic approach just described. But what does it mean to say that the 'commutator corresponds to the Poisson bracket'? Let $f, g \in S$. Suppose we have constructed a deformation of $S[[\hbar]]$ given by a family of bilinear forms B_j, for $j \geq 0$. Forming the commutator of f and g as elements in the ring $S[[\hbar]]$ with this product, we get

$$f \star g - g \star f = \sum_{j \geq 0} \big(B_j(f,g) - B_j(g,f)\big)\hbar^j. \tag{11.1}$$

Since $B_0(f, g) = fg$, the first non-zero term of the power series in (11.1) is $(B_1(f, g) - B_1(g, f))\hbar$. But, according to Dirac, the commutator $f \star g - g \star f$ divided by $i\hbar$ ought to be equal to the Poisson bracket when $\hbar$ goes to zero. Thus $B_1(f, g) - B_1(g, f) = i\{f, g\}$. An easy way to achieve this is to require that $B_1(f, g) = i\{f, g\}/2$, since the Poisson bracket is skew symmetric.

As we saw in §4, the Poisson bracket is really a bidifferential operator in the arguments f and g. Thus we may boldly propose to extend this assumption to all the B_js. The question, of course, is: can one define a $\star$-product in $S[[\hbar]]$ satisfying all these conditions?

The answer is yes. This $\star$-product is called the *Moyal-Weyl product.* It was used by Moyal in [**24**] to study quantum statistical mechanics from the point of view of classical phase space. This product can be described in a very compact way if we use tensor products. First define the differential operator Π: $S \otimes_{\mathbf{C}} S \to S \otimes_{\mathbf{C}} S$ by $\Pi(f \otimes g) = \partial f/\partial p \otimes \partial g/\partial q - \partial g/\partial p \otimes \partial f/\partial q$. Now let Δ: $S \otimes_{\mathbf{C}} S \to S$ be the multiplication map $\Delta(f \otimes g) = fg$. One checks easily that the Poisson bracket can be written using Π and Δ as $\{f, g\} = \Delta\Pi(f \otimes g)$. More generally, the Moyal-Weyl $\star$-product of f and g is $f \star g = \Delta(\exp(i\hbar\Pi)(f \otimes g))$. As an example, let us calculate the coefficient of the term in $\hbar^2$ of $f \star g$. By definition it is $-\Delta(\Pi^2)(f \otimes g)$. An easy calculation shows that this is equal to

$$-\left(\frac{\partial^2 f}{\partial p^2}\frac{\partial^2 g}{\partial q^2} - 2\frac{\partial^2 f}{\partial p \partial q}\frac{\partial^2 g}{\partial p \partial q} + \frac{\partial^2 f}{\partial q^2}\frac{\partial^2 g}{\partial p^2}\right).$$

In particular, if either f or g has degree ≤ 1 then this term is zero.

More generally, if either f or g has degree $\leq k$, then $\Delta(\Pi^{k+1}(f \otimes g)) = 0$. Thus $f \star g$ is indeed a polynomial. Moreover $p \star q - q \star p = i\hbar\{p, q\} = i\hbar$. Normalizing $i\hbar$ to 1 we see that the algebra S with the Moyal-Weyl $\star$-product is isomorphic to $\mathscr{A}$. This is not quite a proof that these two algebras are isomorphic, because we have not verified that the Moyal-Weyl product is associative. We shall not check this here, but leave it, instead, to the conscientious reader as an exercise. For details see [**25**] and [**28**].

The deformation theoretic approach to the Weyl algebra is also interesting from an algorithmic point of view. Calculations with elements of the Weyl algebra are not exactly easy. The multiplication of two monomials of relatively small degree may give rise to a long string of terms. This is awkward to implement in a computer. The $\star$-product approach bypasses all this and gives a closed formula in terms of differentiation of polynomials, a calculation that computers can handle.

12. CONCLUDING REMARKS. G. H. Hardy says in *A Mathematician's Apology* that 'a mathematical idea is 'significant' if it can be connected, in a natural and illuminating way, with a large complex of other mathematical ideas' [**15**, §11]. Having examined the evidence collected in the preceding sections, we can safely say that, by Hardy's criterion, the Weyl algebra is a 'significant idea'. This explains why it has been studied so intensely. A lot is known about the one-dimensional Weyl algebra $\mathscr{A}$. Its right ideals have been classified in [**4**] and [**17**], and its representation theory has been studied very thoroughly [**19**]. The same cannot be said of the many-dimensional Weyl algebras mentioned in §10.

But even $\mathscr{A}$ still hides some secrets. For example, it is not known whether all endomorphisms of $\mathscr{A}$ are surjective. This first appeared in print as 'Problème 11.1' in [**10**]; and it is closely related to the famous *Jacobian conjecture*; see [**6**, Ch. 4, §4].

ACKNOWLEDGMENTS. I wish to thank P. M. Cohn, M. P. Holland, and M. B. Alves for reading a preliminary version of this paper and for their many suggestions and comments. The work on this paper was partially supported by a grant from CNPq (Brazil).

REFERENCES

1. V. I. Arnold, *Mathematical methods of classical mechanics*, second edition, Springer-Verlag, New York, 1989.
2. V. I. Arnol'd and A. B. Givental' *Symplectic geometry*, Dynamical systems IV: symplectic geometry and its applications (V. I. Arnol'd and S. P. Novikov, eds.), Encyclopaedia of Mathematical Sciences, vol. 4, Springer-Verlag, 1990.
3. A. Borel et al., *Algebraic D-modules*, Perspectives in Mathematics 2, Academic Press, Orlando, 1987.
4. R. C. Cannings and M. P. Holland, Right ideals of rings of differential operators, *J. Alg.*, **167** (1994), 116–141.
5. P. M. Cohn, *The Weyl algebra and its field of fractions*, lecture on the occasion of K. A. Hirsch's 80th birthday (unpublished) (1986).
6. S. C. Coutinho, *A primer of algebraic D-modules*, London Mathematical Society Student Texts 33, Cambridge University Press, Cambridge, 1995.
7. P. A. M. Dirac, On quantum algebra, *Roy. Soc. Proc. A*, **110** (1926), 412–418.
8. P. A. M. Dirac, *The principles of quantum mechanics*, Oxford University Press, Oxford, 1958.
9. J. Dixmier, Représentations irrédutibles des algèbres de Lie nilpotentes, *Anais Acad. Bras. Ciênc.* **35** (1963), 491–519.
10. J. Dixmier, Sur les algèbres de Weyl, *Bull. Soc. Math. France*, **96** (1968), 209–242.
11. M. Gerstenhaber, On the deformation of rings and algebras, *Annals of Math.* **79** (1964), 59–103.
12. A. W. Goldie, The structure of prime rings under ascending chain conditions, *Proc. London Math Soc.* **8** (1958) 589–608.
13. A. W. Goldie, Semi-prime rings with maximum condition, *Proc. London Math. Soc.* **10** (1960) 201–220.
14. A. Grothendieck, Eléments de géométrie algébrique IV, Etude locale des schémas et des morphismes de schémas, *Inst. Hautes Etudes Sci. Publ. Math.* **32** (1967).
15. G. H. Hardy, *A mathematician's apology*, Cambridge University Press, Cambridge, 1988.
16. K. A. Hirsch, A note on non-commutative polynomials, *J. London Math. Soc.* **12** (1937) 264–266.
17. L. Le Bruyn, Moduli spaces for right ideals of the Weyl algebra, *J. Alg.* **172** (1995) 32–48.
18. D. E. Littlewood, On the classification of algebras, *Proc. London Math. Soc.* **35** (1933) 200–240.
19. B. Malgrange, *Équations différentielles à coefficients polynomiaux*, Progress in Mathematics 96, Birkhäuser, Boston-Basel-Berlin, 1991.
20. J. C. McConnell and J. C. Robson, *Noncommutative noetherian rings*, John Wiley and Sons, New York, 1987.
21. J. Mehra and H. Rechenberg, *The historical development of quantum theory*, vol. 2: *The discovery of quantum mechanics*, 1925, Springer-Verlag, New York 1982.
22. J. Mehra and H. Rechenberg, *The historical development of quantum theory*, vol. 3: *The formulation of matrix mechanics and its modification*, 1925–1926, Springer-Verlag, New York, 1982.
23. W. Moore, *A life of Erwin Schrödinger*, Cambridge University Press, Cambridge, 1994.
24. J. Moyal, Quantum mechanics as a statistical theory, *Proc. Camb. Phil. Soc.* **45** (1949), 99–124.
25. V. Ovsienko and C. Roger, Deformations of Poisson brackets and extensions of Lie algebras of contact vector fields, *Russian Math. Surveys* **47** (1992), 135–191.
26. I. Percival and D. Richards, *Introduction to dynamics*, Cambridge University Press, Cambridge, 1989.
27. B. L. van der Waerden, *Sources of quantum mechanics*, Dover Publications, New York, 1967.
28. A. Weinstein, Deformation quantization, Séminaire Bourbaki, 46ème année, 1993–94, n^o 789 *Astérisque* **227**, 389–409.

Departmento de Matemática Pura
Instituto de Matemática
Universidade Federal do Rio de Janeiro, P. O. Box 68530
Rio de Janeiro 21945-970, RJ, Brazil
collier@impa.br

Multiple Integrals of Symmetric Functions

Tiberiu Trif

In this paper we illustrate an unitary treatment of a class of multiple Riemann integrals. Let a and b be real numbers such that $a < b$. In the sequel $C_n(a, b)$ denotes the n-dimensional cube $[a, b]^n$, while $D_n(a, b)$ denotes the set of all points $(x_1, \dots, x_n) \in \mathbf{R}^n$ such that $a \le x_1 \le \cdots \le x_n \le b$. A function F from $C_n(a, b)$ to $\mathbf{R}$ is called *symmetric* if $F(x_1, \dots, x_n) = F(x_{\sigma(1)}, \dots, x_{\sigma(n)})$ for all $(x_1, \dots, x_n) \in C_n(a, b)$ and all permutations $\sigma \in S_n$, the symmetric group of order n. The main result to be used is:

Theorem. *Let a and b be real numbers such that $a < b$, and let F be a Riemann integrable symmetric function from $C_n(a, b)$ to $\mathbf{R}$ such that for all $i \in \{1, \dots, n\}$ and all $(x_1, \dots, x_n) \in C_n(a,b)$ the function $F(x_1, \dots, x_{i-1} \cdot x_{i+1}, \dots, x_n)\colon [a ,b] \to \mathbf{R}$ is integrable. Then*

$$\int_{C_n(a,b)} F(x_1, \dots, x_n)\, dx_1 \cdots dx_n = n! \int_{D_n(a,b)} F(x_1, \dots, x_n)\, dx_1 \cdots dx_n.$$

Proof: For any given $\sigma \in S_n$ put

$$D_\sigma := \{(x_1, \dots, x_n) \in C_n(a,b) \mid a \le x_{\sigma(1)} \le \cdots \le x_{\sigma(n)} \le b\},$$

$$I_\sigma := \int_{D_\sigma} F(x_1, \dots, x_n)\, dx_1 \cdots dx_n.$$

Let e be the identity permutation in S_n. Taking into account that F is symmetric and applying Fubini's theorem we obtain

$$\begin{aligned} I_\sigma &= \int_a^b dx_{\sigma(n)} \int_a^{x_{\sigma(n)}} dx_{\sigma(n-1)} \cdots \int_a^{x_{\sigma(2)}} F(x_1, x_2, \dots, x_n)\, dx_{\sigma(1)} \\ &= \int_a^b dx_{\sigma(n)} \int_a^{x_{\sigma(n)}} dx_{\sigma(n\text{-}1)} \cdots \int_a^{x_{\sigma(2)}} F(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)})\, dx_{\sigma(1)} = I_e \end{aligned}$$

for all $\sigma \in S_n$. Hence we have

$$\begin{aligned} &n! \int_{D_n(a,b)} F(x_1, \dots, x_n)\, dx_1 \cdots dx_n \\ &\qquad = n! I_e = \sum_{\sigma \in S_n} I_\sigma = \int_{C_n(a,b)} F(x_1, \dots, x_n)\, dx_1 \cdots dx_n. \qquad \blacksquare \end{aligned}$$

Example 1 [4, Problem 67, p. 61]. Let $f\colon [a, b] \to \mathbf{R}$ be an integrable function. Then

$$\int_{D_n(a,b)} f(x_1) \cdots f(x_n)\, dx_1 \cdots dx_n = \frac{1}{n!} \left[\int_a^b f(x)\, dx \right]^n.$$

Solution: Apply the theorem to $F\colon C_n(a, b) \to \mathbf{R}$ defined by $F(x_1, \dots, x_n) := f(x_1) \cdots f(x_n)$.

Example 2 [**4**, Problem 61, p. 60].

$$\iint_D \ln|\sin(x-y)|\,dx\,dy = -\frac{\pi^2}{2}\ln 2,$$

where $D := \{(x, y) \in \mathbf{R}^2 \,|\, 0 \le x \le y \le \pi\}$.

Solution: The theorem gives

$$\iint_D \ln|\sin(x-y)|\,dx\,dy = \frac{1}{2}I,$$

where

$$I := \int_0^{\pi}\int_0^{\pi}\ln|\sin(x-y)|\,dx\,dy = \int_0^{\frac{\pi}{2}}\int_0^{\frac{\pi}{2}} + \int_0^{\frac{\pi}{2}}\int_{\frac{\pi}{2}}^{\pi} + \int_{\frac{\pi}{2}}^{\pi}\int_0^{\frac{\pi}{2}} + \int_{\frac{\pi}{2}}^{\pi}\int_{\frac{\pi}{2}}^{\pi}.$$

In the second integral we make the substitution $x = u$, $y = \pi/2 + v$, in the third the substitution $x = \pi/2 + u$, $y = v$, and in the fourth the substitution $x = \pi/2 + u$, $y = \pi/2 + v$. Thus we obtain

$$\begin{aligned} I &= 2\int_0^{\frac{\pi}{2}}\int_0^{\frac{\pi}{2}}\ln|\sin(u-v)|\,du\,dv + 2\int_0^{\frac{\pi}{2}}\int_0^{\frac{\pi}{2}}\ln|\cos(u-v)|\,du\,dv \\ &= 2\int_0^{\frac{\pi}{2}}\int_0^{\frac{\pi}{2}}\ln\frac{|\sin(2u-2v)|}{2}\,du\,dv = 2\int_0^{\frac{\pi}{2}}\int_0^{\frac{\pi}{2}}\ln|\sin(2u-2v)|\,du\,dv - \frac{\pi^2}{2}\ln 2. \end{aligned}$$

Using the substitution $s = 2u$, $t = 2v$ we obtain $I = I/2 - (\pi^2\ln 2)/2$. Hence $I = -\pi^2\ln 2$.

Example 3 [**3**, Problem 2.1, p. 49]. Evaluate

$$\int_0^{\infty}\int_0^{\infty}|\ln x - \ln y|e^{-(x+y)}\,dx\,dy.$$

Solution: Let I denote the value of the integral, so $I = \lim_{a\to\infty}I(a)$, where

$$I(a) := \int_0^a\int_0^a|\ln x - \ln y|e^{-(x+y)}\,dx\,dy.$$

The theorem gives

$$\begin{aligned} I(a) &= 2\int_0^a\left\{\int_0^x(\ln x - \ln y)e^{-(x+y)}\,dy\right\}dx \\ &= 2\int_0^a\left\{e^{-x}(1-e^{-x})\ln x - e^{-x}\int_0^x e^{-y}\ln y\,dy\right\}dx \\ &= 2\int_0^a e^{-x}(1-e^{-x})\ln x\,dx - 2\int_0^a e^{-x}\left\{\int_0^x e^{-y}\ln y\,dy\right\}dx. \end{aligned}$$

Integrating by parts in the second integral we get

$$I(a) = 2\int_0^a e^{-x}\ln x\,dx - 4\int_0^a e^{-2x}\ln x\,dx + 2e^{-a}\int_0^a e^{-x}\ln x\,dx.$$

Since

$$\lim_{a\to\infty} e^{-a}\int_0^a e^{-x}\ln x\,dx = 0,$$

we obtain

$$I = 2\int_0^\infty e^{-x}\ln x\,dx - 4\int_0^\infty e^{-2x}\ln x\,dx.$$

Making the substitution $2x = t$ in the second integral we get

$$I = 2\int_0^\infty e^{-x}\ln x\,dx - 2\int_0^\infty e^{-t}\ln\frac{t}{2}\,dt = 2\ln 2.$$

Example 4 [1, p. 27].

$$\int_0^1 \cdots \int_0^1 M_r(x_1,\ldots,x_n)\,dx_1\cdots dx_n = \frac{r}{n+1}$$

where $M_r(x_1,\ldots,x_n)$ denotes the rth largest of $x_1,\ldots,x_n$ (starting with the smallest).

Solution: We have

$$\int_{D_n(0,1)} x_r\,dx_1\cdots dx_n = \int_0^1\int_0^{x_n}\cdots\int_0^{x_r}\int_0^{x_{r-1}}\cdots\int_0^{x_2} x_r\,dx_1\cdots dx_n$$

$$= \frac{1}{(r-1)!}\int_0^1\int_0^{x_n}\cdots\int_0^{x_{r+1}} x_r^r\,dx_r\,dx_{r+1}\cdots dx_n = \frac{r}{(n+1)!}.$$

But, on the other hand, the theorem ensures that

$$\int_0^1 \cdots \int_0^1 M_r(x_1,\ldots,x_n)\,dx_1\cdots dx_n = n!\int_{D_n(0,1)} x_r\,dx_1\cdots dx_n = \frac{r}{n+1}.$$

Remark. In the special cases $r = 1$ and $r = n$ we obtain

$$\int_0^1 \cdots \int_0^1 \min(x_1,\ldots,x_n)\,dx_1\cdots dx_n = \frac{1}{n+1}$$

and

$$\int_0^1 \cdots \int_0^1 \max(x_1,\ldots,x_n)\,dx_1\cdots dx_n = \frac{n}{n+1},$$

respectively.

Example 5. Let $a_1, a_2, \ldots, a_n$ be positive real numbers, and let $a := a_1a_2\cdots a_n$. Then

$$\int_0^{a_1}\int_0^{a_2}\cdots\int_0^{a_n} e^{\max\{a_2^2a_3^2\cdots a_n^2x_1^2,\, a_1^2a_3^2\cdots a_n^2x_2^2,\,\ldots,\, a_1^2a_2^2\cdots a_{n-1}^2x_n^2\}}\,dx_n\cdots dx_2\,dx_1$$

$$= \frac{n}{a^{n-1}}\int_0^a x^{n-1}e^{x^2}\,dx.$$

Solution: Making the substitution $a_2a_3\cdots a_nx_1 = y_1$, $a_1a_3\cdots a_nx_2 = y_2, \dots,$ $a_1a_2\cdots a_{n-1}x_n = y_n$ we obtain

$$\int_0^{a_1}\int_0^{a_2}\cdots\int_0^{a_n} e^{\max\{a_2^2a_3^2\cdots a_n^2x_1^2,\, a_1^2a_3^2\cdots a_n^2x_2^2,\dots,\, a_1^2a_2^2\cdots a_{n-1}^2x_n^2\}}\, dx_n\cdots dx_2\,dx_1$$

$$= \frac{1}{a^{n-1}}\int_{C_n(0,a)} e^{\max\{y_1^2,\dots,y_n^2\}} dy_1\cdots dy_n = \frac{n!}{a^{n-1}}\int_{D_n(0,a)} e^{y_n^2}\, dy_1\cdots dy_n$$

$$= \frac{n!}{a^{n-1}}\int_0^a\int_0^{y_n}\cdots\int_0^{y_2} e^{y_n^2}\, dy_1\cdots dy_n = \frac{n!}{a^{n-1}}\cdot\frac{1}{(n-1)!}\int_0^a y_n^{n-1}e^{y_n^2}\, dy_n.$$

Remark. In the special case $n = 2$ we get

$$\int_0^a\int_0^b e^{\max\{b^2x^2,\, a^2y^2\}}\, dy\,dx = \frac{2}{ab}\int_0^{ab} xe^{x^2}\, dx = \frac{e^{a^2b^2}-1}{ab}.$$

This was a problem in the 50th William Lowell Putnam Competition, 1989 **[2]**.

Example 6 [**5**, p. 893]. Let n be a positive integer, $n \ge 2$, and let $b \in]0,1]$. Then

$$\int_0^1\cdots\int_0^1 \min\left\{1, \frac{b}{x_1},\dots,\frac{b}{x_n}\right\} dx_1\cdots dx_n = \frac{nb-b^n}{n-1}.$$

Solution:

$$\int_{C_n(0,1)} \min\left\{1, \frac{b}{x_1},\dots,\frac{b}{x_n}\right\} dx_1\cdots dx_n$$

$$= n!\int_{D_n(0,1)} \min\left\{1, \frac{b}{x_1},\dots,\frac{b}{x_n}\right\} dx_1\,\cdots\, dx_n = n!\int_0^1\int_0^{x_n}\cdots\int_0^{x_2} \min\left\{1, \frac{b}{x_n}\right\} dx_1\cdots dx_n$$

$$= \frac{n!}{(n-1)!}\int_0^1 x_n^{n-1}\min\left\{1, \frac{b}{x_n}\right\} dx_n = n\int_0^b x^{n-1}\, dx + nb\int_b^1 x^{n-2}\, dx = \frac{nb-b^n}{n-1}.$$

REFERENCES

1. M. S. Klamkin and D. J. Newmann, Inequalities and Identities for Sums and Integrals, *Amer. Math. Monthly* **83** (1976), 26–30.
2. L. F. Klosinski, G. L. Alexanderson, and L. C. Larson, The Fiftieth William Lowell Putnam Mathematical Competition, *Amer. Math. Monthly* **98** (1991), 319–327.
3. B. M. Makarov, M. G. Goluzina, A. A. Lodkin, and A. N. Podkorytov, *Selected Problems in Real Analysis*, Transl. of Math. Monographs, Vol. 107, Amer. Math. Society, Providence, Rhode Island, 1992.
4. G. Pólya and G. Szegő, *Problems and Theorems in Analysis I*, Springer-Verlag, Berlin-Heidelberg-New York, 1972.
5. K. B. Stolarsky, From Wythoff's Nim to Chebyshev's Inequality, *Amer. Math. Monthly* **98** (1991), 889–900.

Universitatea Babeş-Bolyai
Facultatea de Matematică şi Informatică
RO-3400 Cluj-Napoca, Str. Kogălniceanu Nr. 1
Romania
ttrif@math.ubbcluj.ro

Some Inequalities for Principal Submatrices

John Chollet

1. INTRODUCTION. Let A be in $M_n(C)$, the set of all n-by-n complex matrices, and let ω be a nonempty subset of $\{1, 2, \ldots, n\}$ with its elements listed in increasing numerical order. Denote by $A[\omega]$ the principal submatrix of A whose entries are in the intersection of those rows and columns of A specified by ω. If the matrix A is positive definite Hermitian, then so is $A[\omega]$. A known matrix inequality that holds for all positive definite A and all ω is ([**6**], [**7**], [**8**], [**13**, Theorem 7.7.8])

$$A^{-1}[\omega] \geq A[\omega]^{-1}, \tag{1}$$

where the inequality denotes the positive semidefinite ordering (Loewner ordering) of pairs of Hermitian matrices A and B in which $A \geq B$ if $A - B$ is positive semidefinite [**13**, Def. 7.7.1].

In this article, we generalize inequality (1) in two consecutive steps. The first of these is the substitution of a primary matrix function for the matrix inverse appearing in (1). In Section 2, we recall the definition of primary matrix functions, give examples, and obtain a general property involving matrix monotonicity. In Section 3, we discuss matrix convexity, obtain two families of inequalities, and survey some results of Chandler Davis. One of his results states that matrix convexity of all orders is a necessary and sufficient condition on a primary matrix function so that it may replace the matrix inverse in (1). Weakening the condition to convexity of a fixed order leads to an open question posed by Davis.

The next step in the generalization of (1) is to replace the idea of "extracting a principal submatrix" with a more general mapping from $M_n(C)$ to $M_k(C)$. In Section 4, we consider normalized positive linear maps on matrices and review some of the work done by Ando, Choi, Davis, and Kadison. Most of their results were stated in the context of C^*-algebras, but our description is matrix-theoretic.

2. PRIMARY MATRIX FUNCTIONS AND MONOTONICITY. If f is a polynomial function in a variable x and A is a square matrix, it is not difficult to give a definition of the matrix function $f(A)$ by simply requiring that A be substituted for x. This is exactly what is done in the Cayley-Hamilton Theorem. Similarly, in the case of an entire analytic function f, its power series can always be used to define $f(A)$. Perhaps the best known matrix function of this type is the exponential.

Even if f is not analytic, a definition of $f(A)$ exists that agrees with the previous ones for appropriate functions and matrices. Assume f is a real- or complex-valued function whose domain contains the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of a diagonalizable matrix A and suppose A is diagonalized as $A = U diag(\lambda_1, \lambda_2, \ldots, \lambda_n) U^{-1}$. Then the *primary matrix function* associated with the *stem function f* is

$$f(A) = U diag(f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n)) U^{-1}. \tag{2}$$

Proving that $f(A)$ is well-defined, despite the non-uniqueness of a diagonalizing factorization of A, is a straightforward application of Lagrange's interpolation

formula [14, pp. 407-408]; the basic property is that $f(SAS^{-1}) = Sf(A)S^{-1}$ for any nonsingular S [**14**, Theorem 6.2.9(c)]. For Hermitian A, the matrix U may be taken to be unitary ($U^{-1} = U^*$, the complex conjugate transpose of U). The notation $f(A)$ is used with the tacit understanding that the eigenvalues of A are in the domain of f.

Since the matrices considered in this paper are Hermitian (and hence diagonalizable), the definition given in (3) is sufficient for our purposes. However, a primary matrix function whose domain is all complex matrices with eigenvalues in a given set can be defined by using the Jordan canonical form and by requiring that its stem function satisfy certain differentiability conditions; this definition reduces to (2) on diagonalizable matrices [**14**, Section 6.2].

It follows immediately from (2) that if f is a real-valued function and A is Hermitian, then $f(A)$ is also Hermitian. For $f(A)$ to be positive semidefinite or positive definite, f must be nonnegative or positive, respectively, on the eigenvalues of A. Now, for $f(x) = x^{-1}$ and any positive definite A, $A^{-1} = f(A)$ and inequality (1) becomes

$$f(A)[\omega] \geq f(A[\omega]). \tag{3}$$

As we shall see, other functions for which inequality (3) holds are x^2 and $x^{-1/2}$, while the reverse inequality holds for $x^{1/2}$. In the next section, we demonstrate that (3) holds *only* if f is a convex matrix function of order n and that if f is a non-constant monotone matrix function of order $n \geq 2$ then inequality (3) for f^{-1} implies the reverse inequality for f. Moreover, if A is a rank one Hermitian or positive semidefinite Hermitian matrix, then the inequality is true for all of the functions $f_\alpha(x) = x^\alpha$, $\alpha \geq 1$, with some restrictions on α in the Hermitian case. The rest of the discussion in this section is devoted to verifying these assertions and to introducing the concept of matrix monotonicity.

We begin with a matrix version of a result of Kadison [**15**, Theorem 1].

Theorem 1. *If A is Hermitian, then $A^2[\omega] \geq A[\omega]^2$.*

Proof: Let P be a permutation matrix such that PAP^* can be partitioned as follows:

$$\begin{bmatrix} A[\omega] & B \\ B^* & C \end{bmatrix}$$

Then

$$PA^2P^* = \begin{bmatrix} A[\omega] & B \\ B^* & C \end{bmatrix}\begin{bmatrix} A[\omega] & B \\ B^* & C \end{bmatrix} = \begin{bmatrix} A[\omega]^2 + BB^* & A[\omega]B + BC \\ B^*A[\omega] + CB^* & BB^* + CC \end{bmatrix}.$$

If ω has k elements and $\omega' = \{1, 2, \dots, k\}$, then, for any n-by-n matrix E, $(PEP^*)[\omega'] = E[\omega]$. Thus, $A^2[\omega] = A[\omega]^2 + BB^* \geq A[\omega]^2$. ■

It is well-known that if A and B are positive semidefinite and $A \geq B$, then $A^{1/2} \geq B^{1/2}$; a very short proof is given in [**8**, p. 199]. Moreover, the positive semidefinite square root of a positive semidefinite matrix is unique. Using these facts, one obtains the following corollary:

Corollary 1. *If A is positive semidefinite Hermitian, then $A[\omega]^{1/2} \geq A^{1/2}[\omega]$.*

Proof: Note that $A[\omega] = (A^{1/2})^2[\omega] \geq (A^{1/2}[\omega])^2$. Taking square roots, we find that $A[\omega]^{1/2} \geq ((A^{1/2}[\omega])^2)^{1/2} = A^{1/2}[\omega]$. ■

It is also well-known that if $A \geq B > 0$, then $B^{-1} \geq A^{-1}$ ([**8**, p. 198] or [**13**, Corollary 7.7.4]). Applying this result to the inequality in Corollary 1, one gets $(A^{1/2}[\omega])^{-1} \geq (A[\omega]^{1/2})^{-1}$. Inequality (1) and transitivity conclude the proof of the next corollary.

Corollary 2. *If A is positive definite Hermitian, then $A^{-1/2}[\omega] \geq A[\omega]^{-1/2}$*

In the next section, we obtain two families of inequalities that contain the preceding corollaries as special cases.

Loewner in [**17**] introduced a partial order on the set of Hermitian matrices, defined monotone matrix functions in terms of this ordering, and gave necessary and sufficient conditions for a matrix function to be monotone. A real-valued function f on a real interval I is said to be a *monotone matrix function of order n* on I, [**14**, p. 536] if $f(A) \geq f(B)$ whenever the n-by-n Hermitian matrices A and B have eigenvalues in I and satisfy $A \geq B$. If f is matrix monotone for all n, it is called *operator monotone*. Loewner proved that a primary matrix function is an operator monotone matrix function on I if and only if its stem function f is the restriction to I of an analytic function on the upper half-plane $\{z \in C\colon \operatorname{Im} z > 0\}$ that maps the upper half-plane into itself and has finite boundary values [**14**, p. 541]. This *Loewner mapping criterion* can be used to show that the functions x^{α} for $0 < \alpha \leq 1$, $\log x$, and $-x^{-1}$ are operator monotone on $(0, \infty)$ (problem 20 of Section 6.6 of [**14**]).

Conditions for matrix monotonicity (and convexity) can also be given in terms of suitable generalizations of the divided differences appearing in Newton's interpolating polynomial; since notation varies, we use that of [**14**, Section 6.1] for ease of reference. If f is a function whose domain contains the distinct real or complex numbers $t_1, t_2, \ldots, t_n$, then *Newton's interpolating polynomial* (of degree at most $n-1$) is $p_{n-1}(t) = f(t_1) + \Delta f(t_1, t_2)(t - t_1) + \Delta^2 f(t_1, t_2, t_3)(t - t_1)(t - t_2) + \cdots + \Delta^{n-1} f(t_1, t_2, \ldots, t_n)(t - t_1) \cdots (t - t_{n-1})$, where $\Delta^l f(t_1, t_2, \ldots, t_{l+1}) = \sum_{i=1}^{l+1} f(t_i) \prod_{j=1,\, j \neq i}^{l+1} (t_i - t_j)^{-1}$ for $l = 1, 2, \ldots, n-1$. The divided differences $\Delta^l f(t_1, t_2, \ldots, t_{l+1})$ can still be defined even if the points $t_1, t_2, \ldots, t_n$ are not distinct as long as f is analytic on a simply connected domain with these points in its interior or f is continuously differentiable on an real interval containing these points [**14**, Section 6.1.14]. For a continuously differentiable stem function f, the associated matrix function is a monotone matrix function of order n on I if and only if the n-by-n *Loewner matrix* $[\Delta f(t_i, t_j)]$ is positive semidefinite Hermitian for all $t_1, t_2, \ldots, t_n$ in I [**14**, p. 536].

The stem function of a monotone matrix function of order $n \geq 2$ on $[0, \infty)$ is either strictly increasing, and so invertible, or constant (problem 9 in Section 6.6 of [**14**]). This fact and the proof of Corollary 1 for $f(x) = x^{1/2}$ and $f^{-1}(x) = x^2$ suggest the next theorem.

Theorem 2. *Let f be a monotone matrix function of order $n \geq 2$ on $[0, \infty)$. Suppose f is not a constant function on $[0, \infty)$ and suppose $A \geq 0$. If $f^{-1}(A)[\omega] \geq f^{-1}(A[\omega])$, then $f(A[\omega]) \geq f(A)[\omega]$.*

The Spectral Theorem ([**13**, Theorem 4.1.5] or [**20**, p. 223]) ensures that a Hermitian matrix A of rank one can be written as $A = \lambda y y^*$, where λ is the

nonzero eigenvalue and y is a unit eigenvector corresponding to this eigenvalue. If $f(\lambda)$ is defined, then $f(A) = f(\lambda)yy^*$ for any such y.

Theorem 3. *Let $\alpha \geq 1$ and let A be a rank one Hermitian matrix. The inequality $A^\alpha[\omega] \geq A[\omega]^\alpha$ holds for all ω if either (a) $(-1)^\alpha = 1$ or (b) A is positive semidefinite.*

Proof: If $A = \lambda yy^*$, then $A^\alpha = \lambda^\alpha yy^*$. Let $\omega = \{i_1, i_2, \ldots, i_k\}$. Writing $y[\omega]$ for the k-by-1 column matrix whose rows are the rows numbered $i_1, i_2, \ldots, i_k$ of y, one gets $A^\alpha[\omega] = (\lambda^\alpha yy^*)[\omega] = \lambda^\alpha y[\omega]y^*[\omega]$. If $y[\omega] = 0$, then there is nothing to prove because $A[\omega] = A^\alpha[\omega] = 0$. Otherwise $y[\omega]$ is an eigenvector of $A[\omega]$ corresponding to its only nonzero eigenvalue $\lambda\beta$ with $\beta = y^*[\omega]y[\omega] \leq y^*y = 1$. It is then possible to write $A[\omega] = \lambda\beta(\beta^{-1}y[\omega]y^*[\omega])$ and, by the remark before the theorem, $A[\omega]^\alpha = \lambda^\alpha\beta^\alpha(\beta^{-1}y[\omega]y^*[\omega])$. Now

$$A^\alpha[\omega] - A[\omega]^\alpha = \lambda^\alpha y[\omega]y^*[\omega] - \lambda^\alpha\beta^{\alpha-1}y[\omega]y^*[\omega]$$
$$= \lambda^\alpha(1 - \beta^{\alpha-1})y[\omega]y^*[\omega],$$

which is positive semidefinite in either case (a) or (b) because $\beta^{\alpha-1} \leq 1$. ∎

Writing $f^{-1}(x) = x^\alpha$ and $f(x) = x^{1/\alpha}$, one obtains the following as a corollary of Theorems 2 and 3 and the fact that $x^{1/\alpha}$ is a monotone matrix function for $\alpha \geq 1$.

Corollary 3. *If A is a rank one positive semidefinite Hermitian matrix and $\alpha \geq 1$, then $A[\omega]^{1/\alpha} \geq A^{1/\alpha}[\omega]$ for all ω.*

3. CONVEX AND MATRIX CONVEX FUNCTIONS. A real-valued function f defined on an interval I of real numbers is *convex* on I if for all $\lambda_1, \lambda_2, \ldots, \lambda_n$ in I, the inequality $f(u_1\lambda_1 + u_2\lambda_2 + \cdots + u_n\lambda_n) \leq u_1 f(\lambda_1) + u_2 f(\lambda_2) + \cdots + u_n f(\lambda_n)$ holds for all $u_1, u_2, \ldots, u_n$ in $[0,1]$ with $u_1 + u_2 + \cdots + u_n = 1$. Sometimes, for precision, we refer to this type of convexity as *ordinary convexity*. If the second derivative of f exists on a nonempty open interval I, then the nonnegativity of this derivative is necessary and sufficient for f to be convex on I.

Theorem 4. *Let $n > 1$ be a positive integer and f a real-valued function on a real interval I. Fix $i \in \{1, 2, \ldots, n\}$ and set $\omega = \{i\}$. Then $f(A)[\omega] \geq f(A[\omega])$ for every n-by-n Hermitian matrix A whose eigenvalues lie in I if and only if f is convex.*

Proof: Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ in I and let $u_1, u_2, \ldots, u_n$ in $[0,1]$ be such that $u_1 + u_2 + \cdots + u_n = 1$. Construct a unitary matrix U whose ith row is $[\sqrt{u_1}, \sqrt{u_2}, \ldots, \sqrt{u_n}]$ and consider the Hermitian matrix $A = U\mathit{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)U^*$. Now, $f(A)[\omega] = \sum_{j=1}^n f(\lambda_j)u_j$ and $f(A[\omega]) = f(\sum_{j=1}^n \lambda_j u_j)$. Therefore, the inequality in the theorem implies the convexity of f on I. Conversely, if f is convex on I and A is Hermitian with all its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ in I, then $A = U\mathit{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)U^*$ for some unitary $U = [u_{ij}]$. Then $A[w] = \sum_{j=1}^n \lambda_j |u_{ij}|^2$ and $f(A)[\omega] = \sum_{j=1}^n f(\lambda_j)|u_{ij}|^2$. But $\sum_{j=1}^n |u_{ij}|^2 = 1$, so $f(A)[\omega] \geq f(\sum_{j=1}^n \lambda_j |u_{ij}|^2) = f(A[\omega])$. ∎

But is ordinary convexity of f sufficient for inequality (3) to hold for all ω? The following example shows that this is definitely not the case.

Example 1. Consider [**6**, p. 568]

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Then A is positive semidefinite. The function $f(x) = x^4$ is convex. With $\omega = \{1, 2\}$,

$$A^4[w] - A[w]^4 = \begin{bmatrix} 8 & 5 \\ 5 & 3 \end{bmatrix},$$

which is not positive semidefinite.

To give a sufficient condition for inequality (3) to hold, we need to introduce convex matrix functions [**14**, pp. 543-544]. If A and B are n-by-n Hermitian matrices with eigenvalues in an interval I and t is in $[0, 1]$, then $(1 - t)A + tB$ is also Hermitian with eigenvalues in the same interval I. A real-valued function f on I is a *convex matrix function of order n on I* if $(1 - t)f(A) + tf(B) \geq f((1 - t)A + tB)$ for all such A and B and all t in $[0, 1]$; the function f is called *operator convex on I* if it is a convex matrix function of order n on I for all n. Note that matrix convexity of order 1 is ordinary convexity. Concavity of matrix functions is defined by reversing the inequality; f is concave if and only if $-f$ is convex.

Loewner's student, F. Kraus, initiated the study of convex matrix function and proved the following result ([**16**] or [**14**, p. 547]). If f is twice continuously differentiable on I, then f is operator convex on I if and only if the matrix $[\Delta^2 f(t_i, t_j, t_1)]$ is positive semidefinite for all $t_1, t_2, \dots, t_n$ in I. Examples of operator convex matrix functions are x^2, $-x^{1/2}$, $x^{-1/2}$ and x^{-1}; e^x is matrix convex only of order 1 [**14**, Section 6.6, Problem 31, p. 555].

Matrix convexity and monotonicity are related by a theorem of Bendat and Sherman ([**3**, Theorem 3.3] or [**14**, p. 547]): A twice continuous differentiable function f on I is operator convex on I if and only if $\Delta f(t, t_0)$ is operator monotone on I for all t_0 in I. This criterion can be used to prove that the functions x^α for $1 < \alpha < 2$ or for $-1 < \alpha < 0$ are operator convex on $[0, \infty)$ or $(0, \infty)$, respectively. Indeed, the proofs are similar to the one showing the functions $-x^\alpha$ for $0 < \alpha < 1$ have the same property (see problem 17 of Section 6.6 of [**14**] or Corollary 4.1 of [**1**]).

Some relatively recent results relate operator monotonicity, convexity, and concavity in an interesting way. Using Loewner's integral representation [**14**, p. 542] of operator monotone functions, Ando [**1**, Theorem 4] proved that an operator monotone function on $(0, \infty)$ is also operator concave there; see [**11**, Theorem 2] for a different approach to this result. Finally, Mathias [**18**, Theorem 2.1] generalized this result by noticing that the proof in [**11**] shows that any matrix monotone function of order n on $(0, \infty)$ is also matrix concave of order $[n/2]$ there.

Convex matrix functions of order n and inequality (3) are linked in a result of Chandler Davis [8, Theorem 5].

Theorem 5 (Davis). *If f is a convex matrix function of order n on a given interval I, then $f(A)[\omega] \geq f(A[\omega])$ for all $\omega \subseteq \{1, 2, \dots, n\}$ and all n-by-n Hermitian matrices A with eigenvalues in I.*

Davis actually showed that the hypotheses of this theorem imply $Pf(A)P \geq Pf(PAP)P$ for any n-by-n Hermitian projection P and any n-by-n Hermitian matrix A with eigenvalues in I. Davis called this the *Sherman condition on f of order n*. Although this condition seems to be stronger than that of Theorem 5, it is not.

Lemma 1. *Let A be a given n-by-n Hermitian matrix with eigenvalues in an interval I. Then $f(A)[\omega] \geq f(A[\omega])$ for all $\omega \subseteq \{1,2,\ldots,n\}$ if and only if $Pf(A)P \geq Pf(PAP)P$ for any n-by-n Hermitian projection P.*

Proof: If $\omega = \{i_1, i_2, \ldots, i_k\}$ and P is the n-by-n matrix whose i_j column has a 1 in row i_j for $j = 1,2,\ldots,k$ and all other entries are zero, then P is a Hermitian projection and PAP has $A[\omega]$ in the intersection of the rows and columns numbered by ω and zeros everywhere else. Therefore, the Sherman condition implies inequality $f(A)[\omega] \geq f(A[\omega])$ for all $\omega \subseteq \{1,2,\ldots,n\}$. Now assume that this inequality holds for all ω and recall that $V^*f(A)V = f(V^*AV)$ for all unitary matrices V. Let P be any Hermitian projection and V a unitary matrix such that $V^*PV = L = I_k \oplus 0_{n-k}$, the direct sum of the k-by-k identity matrix and the $(n-k)$-by-$(n-k)$ zero matrix. Let $\omega = \{1,2,\ldots,k\}$. Then

$$\begin{aligned}
Pf(A)P &= VLV^*f(A)VLV^* = VLf(V^*AV)LV^* \\
&= V((f(V^*AV)[\omega]) \oplus 0_{n-k})V^* \\
&\geq V(f((V^*AV)[\omega]) \oplus 0_{n-k})V^* = VLf(LV^*AVL)LV^* \\
&= VLf(V^*PAPV)LV^* = VLV^*f(PAP)VLV^* = Pf(PAP)P. \qquad \blacksquare
\end{aligned}$$

Using Theorem 5 and operator convexity of the functions $f(x) = x^\alpha$, we can obtain two families of inequalities for principal submatrices that include some already considered.

Theorem 6. *The inequality $A^\alpha[\omega] \geq A[\omega]^\alpha$ holds for all ω if either (a) $-1 \leq \alpha \leq 0$ and A is positive definite or (b) $1 \leq \alpha \leq 2$ and A is positive semidefinite. The inequality $A[\omega]^\alpha \geq A^\alpha[\omega]$ holds for $0 \leq \alpha \leq 1$ if A is positive semidefinite.*

Part (a) is implied by the second matrix inequality by taking inverses and using inequality (1). Also, the second inequality for $\frac{1}{2} \leq \alpha \leq 1$ can be obtained from part (b) by using monotonicity and Theorem 2.

By Theorem 3, the Sherman condition of order n is satisfied by the function x^3 on the set of positive semidefinite Hermitian matrices of rank one, but this function is not a convex matrix function for any $n \geq 2$ (see problem 30 of Section 6.6 of **[14]**). For $n > 1$, is there a matrix function that satisfies the Sherman condition of order n but is not matrix convex of this order? This is part of a question posed by Davis **[8**, p. 197**]**. His question is also related to the following theorem from **[8]**:

Theorem 7 (Davis). *If f satisfies the Sherman condition of order $2n$, then f is a convex matrix function of order n.*

For $n > 1$, it is not known if there is a function that is matrix convex of order n but does not satisfy the Sherman condition of order $2n$. Similarly, necessary and sufficient conditions on a matrix function for inequality (3) to hold for a fixed n and all ω have not yet been found.

If matrix convexity of a fixed order is strengthened to operator convexity, then Davis **[9]** obtained the following theorem.

Theorem 8 (Davis). *A matrix function f is operator convex on an interval I if and only if f satisfies the Sherman condition of order n for all n.*

4. NORMALIZED POSITIVE LINEAR MAPS. Consider the map $\Phi_\omega(A): M_n(C) \to M_k(C)$ defined by $\Phi_\omega = A[\omega]$. This map has three obvious properties: (1) it is linear; (2) if A is positive semidefinite Hermitian, then so is $\Phi_\omega(A)$; and (3) $\Phi_\omega(I_n) = I_k$, where I_n and I_k are the n-by-n and k-by-k identities. Any map $\Phi: M_n(C) \to M_k(C)$ with these three properties is called a *normalized positive linear map.* Linearity of Φ implies that if $A \geq B$ then $\Phi(A) \geq \Phi(B)$ and that Φ maps Hermitian matrices to Hermitian matrices (write a Hermitian A as the difference of two positive semidefinite matrices). Moreover, if the eigenvalues of the Hermitian matrix A are in (a, b), then so are the eigenvalues of $\Phi(A)$. Indeed, if $\lambda > a$ is the smallest eigenvalue of A, then $A - \lambda I_n \geq 0$ and $\Phi(A - \lambda I_n) = \Phi(A) - \lambda I_k \geq 0$ or $\Phi(A) \geq \lambda I_k$; therefore, the eigenvalues of $\Phi(A)$ are greater than a. A similar calculation with $\gamma I_n - A \geq 0$, where $\gamma < b$ is the largest eigenvalue of A, shows that the eigenvalues of $\Phi(A)$ are smaller than b. In particular, Φ maps positive definite matrices to positive definite matrices

If we replace the extraction of principal submatrices by a normalized positive linear map in (3), we obtain the inequality

$$\Phi(f(A)) \geq f(\Phi(A)). \tag{4}$$

For what primary matrix functions f and for what matrices A does (4) hold for all Φ? In the setting of C^*-algebras, Kadison [**15**] obtained a "Generalized Schwarz Inequality", which shows that (4) holds for $f(x) = x^2$, Hermitian A, and linear maps Φ that preserve the Loewner order and are contractions. Choi [**6**, Theorem 2.1] showed that (4) holds if Φ is a normalized positive linear map, f is operator convex on a symmetric interval $I = (-a, a)$, and A is Hermitian with all eigenvalues in I. It follows easily that the same is true for any interval $I = (a, b)$. This then gives a nice version of Theorem 5 for normalized positive linear maps.

Theorem 9. *If f is an operator convex function on a given interval I and if $\Phi: M_n(C) \to M_k(C)$ is a normalized positive linear map, then $\Phi(f(A)) \geq f(\Phi(A))$ for all n-by-n Hermitian A with eigenvalues in I.*

The next theorem gives two important special cases of Choi's results; the proof is due to Ando [**2**].

Theorem 10. *Let $\Phi: M_n(C) \to M_k(C)$ be any normalized positive linear map and let A in $M_n(C)$ be Hermitian. Then*

$$\Phi(A^2) \geq \Phi(A)^2 \tag{5}$$

and

$$\Phi(A^{-1}) \geq \Phi(A)^{-1} \tag{6}$$

if A is positive definite.

Proof: We begin with the proof of (6). If C and F are positive definite, it is well-known that

$$\begin{bmatrix} C & D \\ D^* & F \end{bmatrix} \geq 0 \text{ if and only if } F \geq D^* C^{-1} D \tag{7}$$

([**1**, Lemma 1] or [**13**, Theorem 7.7.6]). If $B > 0$ and λ in $(0, \infty)$ are given, it follows that

$$\begin{bmatrix} \lambda B & B \\ B & \frac{1}{\lambda} B \end{bmatrix} \geq 0, \tag{8}$$

because $\lambda^{-1}B \geq B^*(\lambda B)^{-1}B = \lambda^{-1}B$. By continuity, (8) also holds if $B \geq 0$. Now, let the positive definite matrix A have eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ with corresponding orthonormal eigenvectors $x_1, x_2, \dots, x_n$. Then $A = \sum_{i=1}^n \lambda_i x_i x_i^*$ and $A^{-1} = \sum_{i=1}^n \lambda_i^{-1} x_i x_i^*$ by the Spectral Theorem; each $\lambda_i > 0$, all $x_i x_i^* \geq 0$, and $\sum_{i=1}^n x_i x_i^* = I_n$. Substituting λ_i for λ and the positive semidefinite matrix $\Phi(x_i x_i^*)$ for B in (8) gives the positive semidefinite Hermitian matrix

$$\sum_{i=1}^n \begin{bmatrix} \lambda_i \Phi(x_i x_i^*) & \Phi(x_i x_i^*) \\ \Phi(x_i x_i^*) & \frac{1}{\lambda_i}\Phi(x_i x_i^*) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n \lambda_i \Phi(x_i x_i^*) & \sum_{i=1}^n \Phi(x_i x_i^*) \\ \sum_{i=1}^n \Phi(x_i x_i^*) & \sum_{i=1}^n \frac{1}{\lambda_i}\Phi(x_i x_i^*) \end{bmatrix}$$

$$= \begin{bmatrix} \Phi(\sum_{i=1}^n \lambda_i x_i x_i^*) & \Phi(\sum_{i=1}^n x_i x_i^*) \\ \Phi(\sum_{i=1}^n x_i x_i^*) & \Phi\left(\sum_{i=1}^n \frac{1}{\lambda_i} x_i x_i^*\right) \end{bmatrix} = \begin{bmatrix} \Phi(A) & \Phi(I_n) \\ \Phi(I_n) & \Phi(A^{-1}) \end{bmatrix}$$

$$= \begin{bmatrix} \Phi(A) & I_k \\ I_k & \Phi(A^{-1}) \end{bmatrix}$$

Therefore, (7) ensures that $\Phi(A^{-1}) \geq \Phi(A)^{-1}$.

Kadison's inequality (5) for positive definite A can be proved in the same way by replacing (8) with

$$\begin{bmatrix} B & \lambda B \\ \lambda B & \lambda^2 B \end{bmatrix} \geq 0$$

in the preceding argument. The result is that

$$\sum_{i=1}^n \begin{bmatrix} \Phi(x_i x_i^*) & \lambda_i \Phi(x_i x_i^*) \\ \lambda_i \Phi(x_i x_i^*) & \lambda_i^2 \Phi(x_i x_i^*) \end{bmatrix} = \begin{bmatrix} I_k & \Phi(A) \\ \Phi(A) & \Phi(A^2) \end{bmatrix} \geq 0$$

and hence that $\Phi(A^2) \geq \Phi(A)^* I_k \Phi(A) \geq \Phi(A)^2$, which is (5) under the additional assumption that A is positive definite. If A is only Hermitian, choose $\lambda > 0$ such that $A + \lambda I_n$ is positive definite. The computation $\Phi(A^2) + 2\lambda\Phi(A) + \lambda^2 I_k = \Phi((A + \lambda I_n)^2) \geq \Phi(A + \lambda I_n)^2 = \Phi(A)^2 + 2\lambda\Phi(A) + \lambda^2 I_k$ gives (5). ∎

As a consequence of his result that an operator monotone function f on $(0, \infty)$ is operator concave there, Ando [**1**, Theorem 4] observed that $f(\Phi(A)) \geq \Phi(f(A))$ for all positive definite A and all normalized positive linear maps. Ando's result gives (6) for $f(x) = -x^{-1}$.

Theorem 9 can be used to obtain inequalities for scalars and Hermitian matrices.

Example 2. For an n-by-n matrix A, let $\Phi(A) = \frac{1}{n}tr(A)$, where $tr(A)$ is the trace of A. If $\lambda_1, \lambda_2, \dots, \lambda_n$ are given nonnegative numbers, form $A = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$ and consider the inequality $\Phi(f(A)) \geq f(\Phi(A))$. For $f(x) = x^\alpha$, this is $\lambda_1^\alpha + \lambda_2^\alpha + \cdots + \lambda_n^\alpha \geq n^{1-\alpha}(\lambda_1 + \lambda_2 + \cdots + \lambda_n)^\alpha$, which is valid for $1 \leq \alpha \leq 2$ if the eigenvalues are nonnegative and for $-1 \leq \alpha \leq 0$ if they are positive. Ando's observation shows that the inequality is reversed for nonnegative eigenvalues and $0 \leq \alpha \leq 1$.

Example 3. Let $\Phi: M_{n^2}(C) \to M_n(C)$ be defined [1, p. 232] as $\Phi\left(\begin{bmatrix} P & Q \\ R & S \end{bmatrix}\right) = \frac{1}{2}(P + S)$, where P, Q, R, and S are in $M_n(C)$. The function $f(x) = \log x$ is operator monotone on $(0, \infty)$ and by Ando's result is also operator concave there. If $C = A \oplus B$, where A and B are n-by-n positive definite Hermitian matrices, then

$\log((A+B)/2) = \log \Phi(C) \geq \Phi(\log C) = (\log A + \log B)/2$. This inequality can be rewritten as $\log((A+B)/2) \geq \log A^{\frac{1}{2}} + \log B^{\frac{1}{2}}$. The same Φ and the matrix $A = \mathrm{diag}(1, -.5)$ can be used to show that inequality (6) does not hold for arbitrary nonsingular Hermitian matrices.

A note on further reading. A complete up-to-date treatment of matrix functions can be found in Chapter 6 of **[14]**. A more elementary presentation with many examples is in **[4]**. For many years, a useful reference has been **[12]**. A historical exposition of the various definitions of the concept of a matrix function is **[19]**. Loewner's theory of monotone matrix functions is treated in detail in **[10]**. Other examples of matrix inequalities obtained from $\Phi(f(A)) \geq f(\Phi(A))$ are in **[1]**.

REFERENCES

1. T. Ando, Concavity of Certain Maps on Positive Definite Matrices and Applications to Hadamard Products, *Linear Algebra and Appl.* **26** (1979), 203–241.
2. T. Ando, communication from Roger A. Horn.
3. J. Bendat and S. Sherman, Monotone and Convex Operator Functions, *Trans. Amer. Math. Soc.* **79** (1955), 58–71.
4. R. Bronson, *Matrix Methods: An Introduction*, Academic Press, New York, 1969.
5. Chi-Kwong Li and R. Mathias, Matrix Inequalities Involving a Positive Linear Map, *Linear and Multilinear Algebra* **41** (1996), 221–231.
6. M. Choi, A Schwarz Inequality for Positive Linear Maps on C^*-Algebras, *Illinois J. Math.* **18** (1974), 565–574.
7. J. Chollet, On Principal Submatrices, *Linear and Multilinear Algebra* **11** (1982), 283–285.
8. C. Davis, Notions Generalizing Convexity for Functions Defined on Spaces of Matrices, in *Convexity: Proceedings of Symposia in Pure Mathematics*, Vol. VII, American Mathematical Society, Providence, R.I., 1963, pp. 187–201.
9. C. Davis, A Schwarz Inequality for Convex Operator Functions, *Proc. Amer. Math. Soc.* **8** (1957), 42–44.
10. W. Donogue, *Monotone Matrix Functions and Analytic Continuation*, Springer, New York, 1974.
11. S. Friedland and M. Katz, On a Matrix Inequality, *Linear Algebra and Appl.* **85** (1987), 185–190.
12. F. R. Gantmacher, *The Theory of Matrices*, Vol. I, Chelsea, New York, 1959.
13. R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.
14. R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.
15. R. V. Kadison, A Generalized Schwarz Inequality and Algebraic Invariants for Operator Algebras, *Ann. of Math.* **56** (1952), 494–503.
16. F. Kraus, Über konvexe Matrixfunctionen, *Math. Zeit.* **41** (1936), 18–42.
17. K. Loewner, Über monotone Matrixfunctionen, *Math. Zeit.* **38** (1934), 177–216.
18. R. Mathias, Concavity of Monotone Matrix Functions of Finite Order, *Linear and Multilinear Algebra* **27** (1990), 129–138.
19. P. C. Rosenbloom, The Equivalence of Definitions of a Matric Function, *Amer. Math. Monthly* **62** (1955), 395–413.
20. G. Strang, *Linear Algebra and Applications*, Academic Press, New York, 1980.

Department of Mathematics
Towson State University
Towson, Maryland 21204
chollet-j@toe.towson.edu

Two Applications of Calculus to Triangular Billiards

Eugene Gutkin

1. INTRODUCTION. The *pedal triangle*, T_1, of a given triangle T is formed by the feet of the three altitudes of T. Let $\mathscr{T}$ be the space of all triangles. Then $p: T \mapsto T_1$ is a natural self-mapping of $\mathscr{T}$, the *pedal mapping*. The dynamics of iterations of p was investigated from several points of view in **[10]**, **[12]**, **[16]**. Here we will study the correspondence $p: T \mapsto T_1$ from yet another, and completely different, perspective.

We assume that the triangle T is acute, and view it as a billiard table. The pedal triangle T_1, which is inscribed in T, is then a periodic billiard orbit (see §2). Moreover, T_1 is the shortest such orbit. Even more remarkable, for the general triangular table, the pedal triangle is the only closed (prime) billiard orbit known!

The first proof, by calculus, that among all inscribed triangles the pedal triangle has the least perimeter, is attributed to J. F. F. Fagnano, ca. 1775. In his honor, the problem just stated is often called the *Fagnano problem* **[4]**, **[5]**, **[15]**. Elementary geometric solutions were later given independently by H. A. Schwarz and L. Fejer **[14]**. Schwarz and Fejer did their work at the end of the 19-th century and in the beginning of the current one. Thus, along with Fagnano, they are the primeval researchers in polygonal billiards! Following tradition, we will call T_1 the *Fagnano orbit*. We reserve the name *Fagnano geodesic* for the Fagnano orbit, traced twice. Indeed, this is a closed geodesic on the degenerate closed Riemannian surface, T_{double}, which is the two-sided T. Every time a geodesic on T_{double} passes by an edge, it changes sides. Thus, if we are tracing T_1 on T_{double}, we have to 'run along it twice' to come back to the starting point.

Let $\mathscr{T}$ be the space of triangles in the euclidean plane, endowed with the natural topology. Two triangles, T and T' are close in this topology if we can label their vertices, $T = ABC$, $T' = A'B'C'$, so that A is close to A', etc. We denote by $\mathscr{A} \subset \mathscr{T}$ the subspace of acute triangles, and think of the elements of $\mathscr{A}$ as triangular billiard tables. The relative length of the Fagnano orbit, $f(T) = |T_1|/|T|$, is a positive continuous function on $\mathscr{A}$. What is the maximum of f on $\mathscr{A}$, and on which $T \in \mathscr{A}$ is this maximum attained? What is the mean value of f with respect to the natural measure on $\mathscr{A}$? In what follows we answer these questions, using elementary calculus.

2. PERIODIC BILLIARD ORBITS. Let $T \subset \mathbf{R}^2$ be a bounded connected region with a piecewise C^1 boundary. The billiard in T is modeled on the motion without friction of a perfectly elastic point-mass (see Fig. 1). For emphasis, we call T the *billiard table*. Equivalently, the billiard in T corresponds to the propagation of light rays inside T, where the boundary of T is a perfect mirror.

Periodic billiard orbits in T correspond to inscribed polygons satisfying the *angle of incidence is equal to the angle of reflection* condition. These are the *harmonic polygons* in the terminology of G. D. Birkhoff **[2]**, or the *light polygons*, according to M. Berger **[1]**.

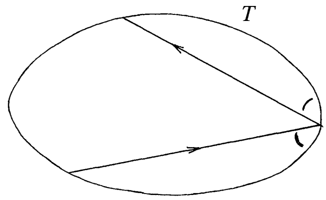


Figure 1. Billiard dynamics in a smooth, convex billiard table.

Let $|P|$ denote the perimeter of a polygon. Then harmonic polygons are the critical points of the function $|P|$ on the space of polygons inscribed in T. This fact is crucial for the theorem about the existence of periodic orbits in any convex C^1 billiard table. Let p, q be positive integers. A periodic orbit with q sides that 'goes p times around the table' has *rotation number* p/q, and *period* q. The theorem asserts that for any rational number, $0 < p/q < 1$, with p and q relatively prime, there are at least two distinct periodic orbits of period q, with rotation number p/q. These are the so-called *Birkhoff periodic orbits* (see **[15]**, **[9]**). This theorem extends, almost verbatim, to a large class of dynamical systems: area preserving twist maps [**9**, chs. 9–13].

If we remove the assumption that the table is C^1, then the preceding assertion breaks down. For instance, no triangular table has a periodic orbit of period 2, with rotation number $1/2$. Obtuse triangles have no periodic orbits of period 3, with rotation number $1/3$.

Little is known about periodic billiard orbits in arbitrary polygons (see **[6]**, **[7]**, **[8]**). It is not even known if every obtuse triangle has a periodic orbit, although there are many partial results in this direction **[4]**.

3. THE FAGNANO PERIODIC ORBIT. In view of the preceding discussion, it is remarkable that the pedal triangle of any acute triangle is a canonical periodic billiard orbit, the Fagnano orbit. Let $T = \Delta ABC$ be an acute triangle with angles A, B, C and sides a, b, c (Fig. 2). Let A_1, B_1, C_1 be the feet of the three altitudes of ABC. The similarity of the four triangles ΔA_1C_1B, ΔA_1B_1C, ΔB_1C_1A, and T yields: angle $AC_1B = C$, angle $C_1A_1B = A$, etc. (see Fig. 2). Thus $T_1 = \Delta A_1B_1C_1$ is a harmonic triangle.

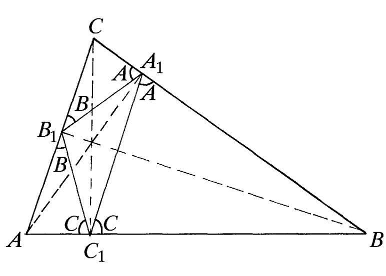


Figure 2. Pedal triangle, or the Fagnano periodic orbit.

Since the pedal triangle T_1 depends continuously on $T \in \mathscr{T}$ in our topology, the length ratio, $|T_1|/|T|$, is a continuous function on $\mathscr{T}$. But the pedal triangle has a billiard interpretation only for acute triangles T. A closely related fact (again, assuming T is acute): T_1 is the unique least-perimeter triangle, inscribed in

T [**1**, ch. 9.4]. In what follows, we treat $T \in \mathscr{A}$ as an independent variable, and call $f(T) = |T_1|/|T|$ the *relative length of the Fagnano orbit*. We regard f as a function on $\mathscr{A}$, thought of as the space of triangular billiard tables.

Theorem 1. *The maximal relative length of the Fagnano periodic orbit is* $1/2$. *It is attained at the equilateral triangles.*

Proof: Let r be the radius of the circle circumscribed about ΔABC. Then [**1**, vol. 1]

$$a = 2r\sin A, \quad b = 2r\sin B, \quad c = 2r\sin C. \tag{1}$$

Similarity of the triangles A_1C_1B, A_1B_1C, B_1C_1A, and ABC yields

$$|A_1B_1| = c\cos C, \quad |B_1C_1| = a\cos A, \quad |A_1C_1| = b\cos B, \tag{2}$$

and hence

$$\frac{|T_1|}{|T|} = \frac{\sin 2A + \sin 2B + \sin 2C}{2(\sin A + \sin B + \sin C)}. \tag{3}$$

From $A + B + C = \pi$, using elementary trigonometry, we obtain

$$\sin 2A + \sin 2B + \sin 2C = 4\sin A \sin B \sin C,$$
$$\sin A + \sin B + \sin C = 4\cos\frac{A}{2}\cos\frac{B}{2}\cos\frac{C}{2}.$$

Thus

$$\frac{|T_1|}{|T|} = 4\sin\frac{A}{2}\sin\frac{B}{2}\sin\frac{C}{2}. \tag{4}$$

Let G be the group of conformal transformations of the plane, i.e., G is generated by isometries and homotheties. The group G acts naturally on $\mathscr{T}$; let $\mathscr{T}/G$ denote the quotient space. Using the angles of a triangle as coordinates on $\mathscr{T}/G$, we identify the closure of $\mathscr{T}/G$ with the simplex $\{(A, B, C)\colon 0 \le A, B, C;\ A + B + C = \pi\}$ in $\mathbf{R}^3$. The projection of $\mathbf{R}^3$ onto the $\{A, B\}$-plane represents $\mathscr{T}/G$ as the triangle $\{(A, B)\colon 0 \le A, B;\ A + B \le \pi\}$. The subspace $\mathscr{A}/G \subset \mathscr{T}/G$ corresponds in this representation to the triangle $\{(A, B)\colon 0 \le A, B \le \pi/2;\ A + B \ge \pi/2\}$; see Fig. 3.

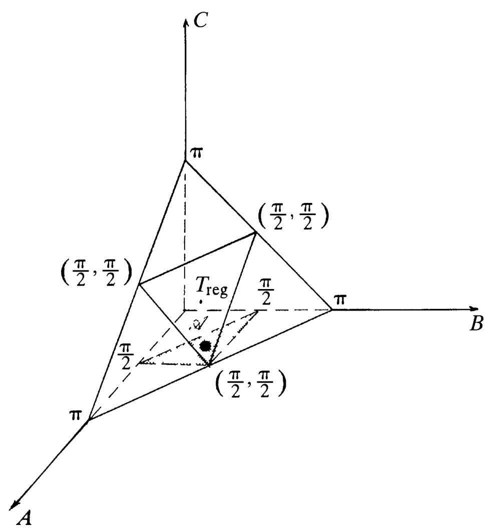


Figure 3. The space of triangles.

Since the billiard dynamics depends only on the intrinsic properties of the table, $\mathscr{T}/G$ is the appropriate space of triangular billiard tables. Setting $x = A/2, y = B/2$, $|T_1|/|T| = f(x, y)$ and using (4), we come to the problem of maximizing the function $f(x, y) = 4 \sin x \sin y \cos(x + y)$ on the closed domain $\mathscr{D} = \{(x, y) : 0 \le x, y \le \pi/4 \le x + y\}$. By our preceding remarks, $\mathscr{D}$ is the closure of $\mathscr{A}/G$, and we think of elements of $\mathscr{D}$ as acute triangular billiard tables.

By elementary trigonometry

$$f_x = 4 \sin y \cos(2x + y), \qquad f_y = 4 \sin x \cos(x + 2y).$$

Hence the only critical point of f in $\mathscr{D}$ is the point $z_0 = (\pi/6, \pi/6) \in int(\mathscr{D})$. A computation shows that the Hessian of f at z_0 is negative definite, so z_0 is a local maximum and the corresponding critical value is $f(z_0) = 1/2$.

The point z_0 corresponds to the equilateral triangle $A = B = C = \pi/3$. In order to conclude that the equilateral triangle has the relatively longest Fagnano orbit, we compute the maximal value of $|T_1|/|T|$ on the boundary, $\partial\mathscr{D}$, which consists of right triangles ABC. By symmetry, it suffices to maximize $|T_1|/|T|$ on the boundary component given by $C = \pi/2$, $0 \le A \le \pi/4$, $B = \pi/2 - A$. This corresponds to the interval $0 \le x \le \pi/8$, $y = \pi/4 - x$ in our parametrization. Restricting $f(x, y)$ to this interval, we obtain $\phi(x) = \sqrt{2}[\cos(\pi/4 - x) - 1/\sqrt{2}]$. This function is monotonically increasing on $[0, \pi/8]$, from $\phi(0) = 0$ to $\phi(\pi/8) = \sqrt{2} - 1$. Thus, the $(\pi/4, \pi/4, \pi/2)$-triangle maximizes $|T_1|/|T|$ on $\partial\mathscr{D}$, and the corresponding maximal value is $\sqrt{2} - 1 < 1/2$. ■

The space $\mathscr{T}$ is, essentially, $\mathbf{R}^6$. The natural measure on $\mathscr{T}$ corresponds to the Lebesgue measure on $\mathbf{R}^6$. Under the projection $\mathscr{T} \to \mathscr{T}/G$ it reduces to the Lebesgue measure $c\,dx\,dy$ on the square $\{(x, y) : 0 \le x, y \le \pi/4\}$. The value of the constant $c > 0$ is irrelevant for our purpose, and we set $c = 1$.

Theorem 2. *The average relative length of the Fagnano periodic orbit, with respect to the Lebesgue measure on the space of acute triangles is* $12\pi^{-1} - 24\pi^{-2} - 1 \simeq .39$.

Proof: We use the notation and the setting of the proof of Theorem 1. By our preceding remarks, the quantity we are after is the mean value, with respect to the Lebesgue measure, of $f(x, y) = 4 \sin x \sin y \cos(x + y)$ on the region $\mathscr{D} = \{(x, y) : 0 \le x, y \le \pi/4 \le x + y\}$.

Set $I = \iint_{\mathscr{D}} f(x, y)\,dx\,dy$. By elementary trigonometry, $f(x, y) = \cos 2x + \cos 2y - \cos 2(x + y) - 1$. Using the symmetry of $\mathscr{D}$, we write

$$I = 2\iint_{\mathscr{D}} \cos 2x\,dx\,dy - \iint_{\mathscr{D}} \cos 2(x + y)\,dx\,dy - area(\mathscr{D}).$$

The integrals are straightforward to evaluate, and we omit the details. We have

$$\iint_{\mathscr{D}} \cos 2x\,dx\,dy = -\iint_{\mathscr{D}} \cos 2(x + y)\,dx\,dy = \frac{1}{4}\left(\frac{\pi}{2} - 1\right).$$

Thus

$$I = \frac{3}{4}\left(\frac{\pi}{2} - 1\right) - area(\mathscr{D}).$$

Since $\mathscr{D}$ is the right isoceles triangle with side $\pi/4$, we have $area(\mathscr{D}) = \pi^2/32$, and the mean value is:

$$\langle f \rangle = \frac{12(\pi - 2)}{\pi^2} - 1. \qquad ■$$

Recall that $f(T)$ is the relative length of the Fagnano periodic orbit in T. By the proof of Theorem 1, the range of f is the interval $[0, .5]$. Note that the mean value of f, which is approximately .39, is much closer to .5 than to zero.

4. CONCLUDING REMARKS. The equilateral triangle almost invariably pops up as the solution to any geometric—analytic optimization problem for triangles **[13]**. Thus, Theorem 1 is far from surprising, and may well be in the literature. We note that the equilateral triangle is the unique fixed point of the pedal mapping, $p: T \mapsto T_1$, of $\mathcal{T}/G$. Let $T \in \mathcal{T}/G$ be arbitrary, and set $T_n = p^n(T)$, the 'pedal triangle of n-th generation'. For almost any T the sequence $T = T_0, T_1, T_2, \dots$ is infinite, i. e., the triangles T_n never degenerate. The equilateral triangle is characterized by the condition that T_n is acute for all n. Both remarks follow from the representation of the pedal mapping in **[12]**.

Theorems 1 and 2 suggest a possible approach to the elusive problem of periodic orbits in polygons [3], **[11]**. Let τ be a specific type of orbits, and let $\mathcal{P} = \mathcal{P}(\tau)$ be the space of polygons that have a periodic orbit of type τ. The billiard orbit of a given type in a polygonal table, if it exists, is essentially unique. Let $P_1 \subset P$ be the corresponding inscribed polygon. The ratio $|P_1|/|P| = f_\tau(P)$ is the relative length of the orbit of type τ. We view f_τ as a function on $\mathcal{P}$. It may be useful to investigate these functions from the point of view of Theorems 1 and 2.

REFERENCES

1. M. Berger, *Geometry*, vols. I, II, Springer-Verlag, Berlin, 1987.
2. G. D. Birkhoff, *Collected Mathematical Papers*, American Mathematical Society, Providence, RI, 1950.
3. H. T. Croft, K. J. Falconer, and R. K. Guy, *Unsolved Problems in Geometry*, Springer-Verlag, New York, 1991.
4. G. Galperin, A. Stepin, and Y. Vorobets, Periodic orbits in polygons; generating mechanisms, *Russian Math. Surveys* 47 (1992), 5–80.
5. G. Galperin and A. Zemlyakov, *Mathematical Billiards*, Nauka, Moscow, 1990.
6. E. Gutkin, Billiards in polygons, *Physica* 19 D (1986), 311–333.
7. E. Gutkin, Billiards in polygons: Survey of recent results, *J. Stat. Phys.* 83 (1996), 7–26.
8. E. Gutkin and S. Troubetzkoy, Directional flows and strong recurrence for polygonal billiards, *Pitman Res. Notes in Math.*, vol. 262, 1996, 21–45.
9. A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, Cambridge, 1995.
10. J. G. Kingston and J. L. Synge, The sequence of pedal triangles, *Amer. Math. Monthly* 95 (1988), 809–822.
11. V. Klee and S. Wagon, *Old and New Unsolved Problems in Plane Geometry and Number Theory*, Mathematical Association of America, 1991.
12. P. Lax, The ergodic character of sequences of pedal triangles, *Amer. Math. Monthly* 97 (1990), 377–381.
13. G. Polya and G. Szego, *Isoperimetric Inequalities in Mathematical Physics*, Princeton University Press, Princeton, 1951.
14. H. Rademacher and O. Toeplitz, *Von Zahlen und Figuren*, Springer-Verlag, Berlin, 1933. (English version: *The enjoyment of mathematics*, Princeton University Press, Princeton, 1957.)
15. S. Tabachnikov, *Billiards*, "Panoramas et Synthèses" n. 1, Société Mathématique de France, Paris, 1995.
16. P. Ungar, Mixing property of the pedal mapping, *Amer. Math. Monthly* 97 (1990), 898–900.

Mathematics Department
University of Southern California
Los Angeles, CA 90089-1113
egutkin@math.usc.edu

Early Transcendentals

Steven H. Weintraub

Several current calculus texts have "early transcendental" versions, in which the exponential and logarithm functions are introduced early in the text. These functions are usually justified by various "hand-waving" arguments. The point of this article is to show how they may be introduced rigorously.

We proceed in two steps. We begin with the basic existence theorem (BET): There is a differentiable function $y = f(x)$, defined for $x \geq 0$, satisfying

$$f(0) = 1, \qquad f'(x) = f(x) \qquad (*)$$

for all $x \geq 0$. From the BET we derive that there is a unique function $f(x)$, defined and differentiable for all real x, that satisfies $(*)$ for all real x. This function is of course the exponential function. We then derive the basic properties of the exponential function. These derivations provide a beautiful illustration of some of the basic elements of calculus: the greatest lower bound property, the intermediate value theorem, the mean value theorem, and the chain rule.

We then prove the BET by using power series. At first glance this may seem to delay the introduction of transcendental functions even more than the usual approach, as power series are introduced rather late in most calculus courses (though, logically speaking, it is not necessary to do so). However, $(*)$ is carefully chosen so that we may prove it with only a minimum of the theory of power series —we need consider only series with all terms non-negative, and the only convergence criterion we use is comparison with a geometric series. Thus this material may be introduced separately, near the beginning of the course, and returned to later, when power series are considered in detail. This approach has another advantage. Introducing the power series for the exponential function early not only introduces the exponential function early, but also provides a method for actually computing it.

1. THE BASIC EXISTENCE THEOREM AND ITS CONSEQUENCES

Theorem 1.1 . (*The basic existence theorem = BET*) *There is a differentiable function $y = f(x)$, defined for $x \geq 0$, satisfying*

$$f(0) = 1, \qquad f'(x) = f(x) \qquad (*)$$

for all $x \geq 0$.

For the remainder of this section we *assume* the BET. We *prove* it in the next section, as Corollary 2.16.

Note that the BET does not tell us that there is such a function $f(x)$, much less a *unique* such function, defined for *all* real x. Our first objective is to derive this from the BET. Actually, we prove a slightly more general result in Theorem 1.4.

Lemma 1.2. *If $f(x)$ is a function satisfying the BET, then $f(x) > 0$ for all $x \geq 0$.*

Proof: Suppose there is an $x \geq 0$ with $f(x) \leq 0$. Then the set $S = \{x \geq 0 \mid f(x) \leq 0\}$ is non-empty. Since S is bounded from below by 0, S has a greatest lower bound $x_0 \geq 0$. Then, by the continuity of f, we have $f(x_0) \leq 0$. Since $f(0) = 1$, we have $x_0 > 0$. Then, $1 = f(0) > f(x_0)$, so, by the mean value theorem, there is an x_1 with $0 < x_1 < x_0$, and such that $f'(x_1) < 0$. Thus $f(x_1) = f'(x_1) < 0$, and so $x_1 \in S$. But $x_1 < x_0$, which contradicts x_0 being a lower bound of S. We conclude that $f(x) > 0$ for all $x \geq 0$. ■

Lemma 1.3. *There is a function $e(x)$ that satisfies $(*)$ for all $x \in \mathbb{R}$, and furthermore has $e(x) > 0$ for all $x \in \mathbb{R}$.*

Proof: Let $f(x)$ be a function satisfying the BET. Define $e(x)$ by

$$e(x) = \begin{cases} f(x) & x \geq 0 \\ 1/f(-x) & x \leq 0. \end{cases}$$

This definition is consistent for $x = 0$ as $f(0) = 1 = 1/f(0)$. Note that, by Lemma 1.2, $e(x)$ is defined and positive for all $x \in \mathbb{R}$, and is always positive. Clearly $e(x)$ is differentiable for $x > 0$. As a composition of differentiable functions, $e(x)$ is differentiable for $x < 0$. Clearly $e'(x) = e(x)$ for $x > 0$. For $x < 0$ we have, by the chain rule,

$$\begin{aligned} \frac{d}{dx}(e(x)) &= \frac{d}{dx}(1/f(-x)) = \left(-f'(-x)/f(-x)^2\right) \cdot (-1) \\ &= f'(-x)/f(-x)^2 = f(-x)/f(-x)^2 = 1/f(-x) = e(x). \end{aligned}$$

For the same reasons, the derivative from the right at $x = 0$ exists and has the value 1, and the derivative from the left at $x = 0$ exists and has value 1, so $e'(0)$ exists and has value 1. Thus $e'(x)$ exists for all x, $e'(0) = 1$, and $e'(x) = e(x)$ for all x, so $e(x)$ satisfies $(*)$ for all $x \in \mathbb{R}$. ■

Theorem 1.4. *Fix real numbers α and b. There exists a unique differentiable function g satisfying*

$$g(0) = b, \qquad g'(x) = \alpha g(x) \qquad (**)$$

for all $x \in \mathbb{R}$.

Proof: Let $e(x)$ be *any* function satisfying Lemma 1.3. Define $g(x)$ by $g(x) = be(\alpha x)$. Then $g(0) = be(0) = b \cdot 1 = b$ and, by the chain rule,

$$g'(x) = b(e'(\alpha x) \cdot \alpha) = \alpha(be'(\alpha x)) = \alpha(be(\alpha x)) = \alpha g(x)$$

so $g(x)$ satisfies $(**)$.

Now let $g(x)$ be *any* function satisfying $(**)$. Let $h(x)$ be the quotient $h(x) = g(x)/e(\alpha x)$. (Note the denominator is never zero.) Then $h(0) = g(0)/e(0) = b/1 = b$ and, again by the chain rule,

$$h'(x) = \frac{e(\alpha x)(\alpha g(x)) - (\alpha e(\alpha x))g(x)}{e(\alpha x)^2} = 0$$

for all x, so $h(x)$ is constant. Hence, $h(x) = h(0) = b$ for all $x \in \mathbb{R}$, i.e., $g(x) = be(\alpha x)$. Choosing $b = 1$ and $\alpha = 1$ shows that $g(x) = e(x)$ in this case, i.e., that there is a unique function satisfying the conclusion of Lemma 1.3. Letting b and α be arbitrary gives Theorem 1.4 in general. ■

Now that we have *unique* functions with our desired properties, we can give them names.

Definition 1.5. For a fixed real number α, let $\exp_\alpha(x)$ be the unique differentiable function satisfying

$$\exp_\alpha(0) = 1, \qquad \exp'_\alpha(x) = \alpha\exp_\alpha(x) \text{ for all } x \in \mathbb{R}.$$

We abbreviate $\exp_1(x)$ by $\exp(x)$.

Corollary 1.6. *For any fixed real number* α, $\exp_\alpha(x) = \exp(\alpha x)$.

Proof: $\exp(\alpha 0) = 1 = \exp_\alpha(0)$, $\exp'(\alpha x) = \alpha\exp(\alpha x)$, and $\exp'_\alpha(x) = \alpha\exp_\alpha(x)$, so, by the uniqueness part of Theorem 1.4, $\exp(\alpha x) = \exp_\alpha(x)$ for all $x \in \mathbb{R}$. ∎

The next two theorems give the basic properties of the exponential function.

Theorem 1.7. *The function* $\exp(x)$ *has the following properties*:

a) $\exp(x) > 0$ *for all* $x \in \mathbb{R}$.
b) $\exp(x)$ *is strictly increasing*.
c) $\exp(x) \geq 1 + x$ *for all* $x \in \mathbb{R}$.
d) $\lim_{x\to+\infty}\exp(x) = +\infty$, $\lim_{x\to-\infty}\exp(x) = 0$.
e) *For any* n, $\lim_{x\to+\infty}\exp(x)/x^n = \infty$.

Proof:

a) $\exp(x) = e(x)$ is positive for all $x \in \mathbb{R}$ by Lemma 1.3.
b) $\exp'(x) = \exp(x) > 0$ for all x so, by the mean value theorem, $\exp(x)$ is strictly increasing.
c) If $x = 0$ then $\exp(x) = 1 + x$. Suppose $x > 0$. Then, by the mean value theorem, $\exp(x) - \exp(0) = \exp'(c)(x - 0)$ for some c with $0 < c < x$. But $\exp'(c) = \exp(c) > \exp(0) = 1$, so $\exp(x) > \exp(0) + \exp(c)x > 1 + x$ for $x > 0$. A similar application of the mean value theorem shows that $\exp(x) > 1 + x$ for $x < 0$.
d) It follows immediately from c) that $\lim_{x\to+\infty}\exp(x) = +\infty$. By the uniqueness of $\exp(x)$, and by Lemma 1.3, $\exp(x) = 1/\exp(-x)$ so $\lim_{x\to-\infty}\exp(x) = 0$.
e) Let $f(x) = \exp(x)/x^{n+1}$. Then $f'(x) = (x^{n+1}\exp(x) - \exp(x)nx^n)/x^{2n+2} = (x-n)\exp(x)/x^{n+2} > 0$ for $x > n$. Hence $f(x) > f(n)$ for $x > n$, i.e., $\exp(x)/x^{n+1} > \exp(n)/n^{n+1}$ for $x > n$, so $\exp(x)/x^n > x \cdot \exp(n)/n^{n+1}$ for $x > n$, and so $\lim_{x\to+\infty}\exp(x)/x^n = +\infty$. ∎

Theorem 1.8. *For any real number* α, *and for any real numbers* w *and* x, $\exp_\alpha(w + x) = \exp_\alpha(w)\exp_\alpha(x)$.

Proof: Fix α and w and set $h(x) = \exp_\alpha(x + w)/\exp_\alpha(w)$. (Note the denominator is non-zero.) Then $h(0) = 1$ and $h'(0) = \alpha h(0)$. Comparing Definition 1.5, we see that $h(x) = \exp_\alpha(x)$, i.e.,

$$\exp_\alpha(x) = \frac{\exp_\alpha(x + w)}{\exp_\alpha(w)}.$$ ∎

Corollary 1.9. *For any integer* n *and all* $x \in \mathbb{R}$, $\exp_\alpha(nx) = \exp_\alpha(x)^n$.

Proof: Induction and Theorem 1.8. ■

Definition 1.10. The real number e is defined by $e = \exp(1)$. (Note $e = \exp(1) > \exp(0) = 1$.)

Theorem 1.8 and Corollary 1.9 justify the notation

$$e^x = \exp(x) \tag{1.11}$$

as $e^0 = 1$, $e^{-1} = 1/e$, $e^{m+n} = e^m \cdot e^n$ and $e^{mn} = (e^m)^n$, the usual laws of exponents, for any rational numbers m and n.

The continuity of $\exp(x)$, and parts b) and d) of Corollary 1.7, imply that the exponential function $\exp : \mathbb{R} \to (0, \infty)$ has a continuous inverse.

Definition 1.12. The *natural logarithm function* $\ln : (0, \infty) \to \mathbb{R}$ is the inverse of $\exp : \mathbb{R} \to (0, \infty)$.

From Definition 1.2, Theorem 1.8 and Corollary 1.9 (with $\alpha = 1$) it is easy to derive the usual properties for logarithms: for $x, y > 0$, $\ln(xy) = \ln(x) + \ln(y)$, $\ln(1/x) = -\ln(x)$, $\ln(x^n) = n\ln(x)$. More interesting are the following properties:

Proposition 1.13

a) $\ln(1) = 0$.
b) $\ln(x)$ *is differentiable for all* $x \in \mathbb{R}$ *and* $\ln'(x) = 1/x$.

Proof: Part a) is immediate from $1 = \exp(0)$. As for part b), recall a general result about inverse functions: If f is a differentiable function with inverse g, and $f'(a) \neq 0$, then g is differentiable at $f(a)$ and $g'(f(a)) = 1/f'(a)$. We apply this here with $f = \exp$ and $g = \ln$. Then $f'(a) = f(a) \neq 0$ for any a. Set $a = \ln(x)$. Then $f(a) = x$, so $g'(x) = g'(f(a)) = 1/f'(a) = 1/f(a) = 1/x$. ■

Remark 1.14. It is common to *define* $\ln(x)$ by

$$\ln(x) = \int_1^x 1/t\,dt. \tag{1.15}$$

If we define $\ln(x)$ by Definition 1.12, then (1.15) becomes a *theorem* that is an immediate application of Proposition 1.13 and the fundamental theorem of calculus.

For any positive number a, we can now define a^x by

$$a^x = \exp_{\ln(a)}(x) = \exp(x\ln(a)). \tag{1.16}$$

Again a^x satisfies the usual laws of exponents ($a^0 = 1, a^1 = a, a^{m+n} = a^m a^n$, $(a^m)^n = a^{mn}$). Also, by Definition 1.5,

$$\frac{d}{dx}(a^x) = \frac{d}{dx}\big(\exp_{\ln(a)}(x)\big) = \ln(a)\exp_{\ln(a)}(x) = a^x\ln(a). \tag{1.17}$$

2. PROOF OF THE BASIC EXISTENCE THEOREM. In this section we need to use some basic facts about series. However, we always deal with series with non-negative terms, and that simplifies the proofs of these facts. For a series with

non-negative terms has increasing partial sums, and it is easy to show that an increasing sequence that is bounded from above converges to its least upper bound. Beyond this, all we need is that if Σb_n and Σc_n are series of non-negative terms that converge to B and C respectively, then $\Sigma b_n + c_n$ converges to $B + C$, and $\Sigma t b_n$ converges to tB for $t \geq 0$. If, in addition, $b_n \geq c_n$ for all n, we need to know that $B \geq C$, and also that $\Sigma(b_n - c_n)$ converges to $B - C$.

We now consider series of the form

$$\sum_{n=0}^{\infty} a_n x^n, \qquad x \in \mathbb{R}.$$

If, for some value of x, this series converges to some value S, we say that the sum of the series is $S = S(x)$. In this way we define a function, and we write

$$S(x) = \sum_{n=0}^{\infty} a_n x^n. \tag{2.1}$$

Definition 2.2. $S(x)$ has non-negative coefficients if $a_n \geq 0$ for all $n \geq 0$.

Definition 2.3. Let $E(x)$ be the series

$$E(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots.$$

We will be interested only in series with non-negative coefficients, and the sums of these series for values of $x \geq 0$; in particular, we are most interested in the series $E(x)$, and its values for $x \geq 0$. In this general situation the partial sums of $S(x)$ always form an increasing sequence.

Definition 2.4. A series $S(x)$ with non-negative coefficients is *eventually geometrically dominated* for $x = x_0$ if there is a real number C, a real number r with $0 \leq r < 1$, and an integer N such that $a_n x_0^n \leq C r^n$ for all $n \geq N$.

Definition 2.5. A series $S(x)$ is called *peaceful* if

a) $S(x)$ has non-negative coefficients, and
b) $S(x)$ is eventually geometrically dominated for every $x \geq 0$.

Every series $S(x)$ is eventually geometrically dominated for $x = 0$ (choose $C = 0$, $r = 0$, and $N = 1$), so the condition in Definition 2.5 b) is non-trivial only for $x > 0$.

Lemma 2.6. *Let* $S(x) = \sum_{n=0}^{\infty} a_n x^n$ *with* $a_n \geq 0$ *for all* n, *and suppose that* $a_n > 0$ *for* n *sufficiently large and that* $\lim_{n\to\infty} a_{n+1}/a_n = 0$. *Then* $S(x)$ *is peaceful.*

Proof: Let $x > 0$ be arbitrary. Choose N so that $a_n > 0$ and $a_{n+1}/a_n < 1/(x+1)$ for $n \geq N$; this is possible as $\lim_{n\to\infty} a_{n+1}/a_n = 0$. Let $r = x/(x+1) < 1$. For any $n \geq N$ we have

$$\begin{aligned} a_n x^n &= \left(x\frac{a_n}{a_{n-1}}\right)\left(x\frac{a_{n-1}}{a_{n-2}}\right)\cdots\left(x\frac{a_{N+1}}{a_N}\right)a_N x^N \\ &\leq a_N x^N r^{n-N} = \{a_N x^N r^{-N}\} r^n, \end{aligned}$$

so $S(x)$ is eventually geometrically dominated. ■

Corollary 2.7. *$E(x)$ is peaceful.*

Proof: For $E(x)$, $a_n = 1/n!$, so $a_{n+1}/a_n = 1/(n+1)$ and $\lim_{n\to\infty} a_{n+1}/a_n = 0$. ∎

Lemma 2.8. *Let $S(x) = \Sigma_{n=0}^{\infty} a_n x^n$ be a peaceful series. Then this series converges for every $x \geq 0$.*

Proof: Fix $x \geq 0$. Since the coefficients a_n are all non-negative, and $x \geq 0$, each term in the series is non-negative, and so the partial sums form an increasing sequence. $S(x)$ is eventually geometrically dominated, so pick C, r, and N as in Definition 2.4. Then

$$\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{N-1} a_n x^n + \sum_{n=N}^{\infty} a_n x^n.$$

Let us compare this with the series

$$\sum_{n=0}^{N-1} a_n x^n + \sum_{n=N}^{\infty} Cr^n.$$

Consider this second series. Its first summand is a finite summand; denote its sum by S_1. Its second summand is a geometric series with sum $Cr^N/(1-r)$; in particular, this sum is an upper bound of the partial sums. Each term in the first series is less than or equal to the corresponding term in the second series, so the partial sums of the first series are bounded from above by $S_1 + Cr^N/(1-r)$. Thus, this series converges. ∎

Lemma 2.9. *Let $S(x)$ be peaceful. Then $S(x)$ is an increasing function on* $[0, \infty)$.

Proof: If $x_2 > x_1$,

$$S(x_2) - S(x_1) = \sum_{n=0}^{\infty} a_n x_2^n - \sum_{n=0}^{\infty} a_n x_1^n = \sum_{n=0}^{\infty} a_n (x_2^n - x_1^n) \geq 0,$$

i.e., $S(x_2) \geq S(x_1)$. ∎

Lemma 2.10. *Let $S(x)$ be peaceful. Then $S(x)$ is continuous for all $x \geq 0$.*

Proof: We have to show that for every $x_0 \geq 0$, and for every $\varepsilon > 0$, there exists a δ such that $|S(x) - S(x_0)| < \varepsilon$ whenever $|x - x_0| < \delta$ and $x \geq 0$. Let $x_1 = x_0 + 1$. Then $S(x_1)$ exists, and is the limit of its partial sums, so there is an N such that for any $x < x_1$,

$$0 \leq \sum_{n=N+1}^{\infty} a_n x^n \leq \sum_{n=N+1}^{\infty} a_n x_1^n = S(x_1) - \sum_{n=0}^{N} a_n x_1^n < \varepsilon/2.$$

Now $f(x) = \Sigma_{n=0}^{N} a_n x_n$ is a polynomial, and hence a continuous function, so there exists a $\delta_1 > 0$ with $0 \leq f(x) - f(x_0) < \varepsilon/2$ for $x_0 \leq x < x + \delta_1$. Thus, if we let $\delta_+ = \min(1, \delta_1)$, then for $x_0 \leq x < x_0 + \delta_+$,

$$\begin{aligned} 0 \leq S(x) - S(x_0) &= \sum_{n=0}^{\infty} a_n x^n - \sum_{n=0}^{\infty} a_n x_0^n \\ &\leq \sum_{n=0}^{\infty} a_n x^n - \sum_{n=0}^{N} a_n x_0^n = \sum_{n=0}^{N} a_n x^n + \sum_{n=N+1}^{\infty} a_n x^n - \sum_{n=0}^{N} a_n x_0^n \\ &= \sum_{n=N+1}^{\infty} a_n x^n + \left(\sum_{n=0}^{N} a_n x^n - \sum_{n=0}^{N} a_n x_0^n \right) < \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

A similar (in fact, easier) argument for $x < x_0$ (taking $x_1 = x_0$ there) shows there is a δ_- with $-\varepsilon < S(x) - S(x_0) \le 0$ for $x_0 - \delta_- < x \le x_0$; setting $\delta = \min(\delta_+, \delta_-)$ we see that $|S(x) - S(x_0)| < \varepsilon$ for $|x - x_0| < \delta$, $x \ge 0$, as required. ∎

Definition 2.11. For a given series $S(x) = \sum_{n=0}^{\infty} a_n x^n$, let $\tilde{S}(x) = \sum_{n=0}^{\infty}(n+1)a_{n+1}x^n$ be the series obtained by differentiating $S(x)$ term-by-term.

Example 2.12. For the series $E(x)$, we have

$$\begin{aligned}\tilde{E}(x) &= 1 + \frac{2x}{2!} + \frac{3x^2}{3!} + \frac{4x^3}{4!} + \cdots \\ &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = E(x).\end{aligned}$$

Lemma 2.13. *Let $S(x) = \sum_{n=0}^{\infty} a_n x^n$ be peaceful. Then $\tilde{S}(x)$ is also peaceful.*

Proof: Write $\tilde{S}(x) = \sum_{n=0}^{\infty} b_n x^n$, so $b_n = (n+1)a_{n+1} \ge 0$ for all n. Pick an $x > 0$. Then, since $S(x)$ is peaceful, there are a real number B, an integer M, and a $q < 1$ with $a_n x^n \le Bq^n$ for all $n \ge M$. Choose $N \ge \max(M, q/(1-q))$. A little algebra shows that if $N \ge q/(1-q)$, $r = q(N+1)/N < 1$. Then, for $n \ge N$,

$$b_n x^n = (n+1)a_{n+1}x^n < (n+1)Bq^{n+1}/x.$$

But for $n \ge N$,

$$n + 1 = \left(\frac{n+1}{n}\right)\left(\frac{n}{n-1}\right)\left(\frac{n-1}{n-2}\right)\cdots\left(\frac{N+1}{N}\right)N$$

and, since

$$\frac{n+1}{n} < \frac{n}{n-1} < \cdots < \frac{N+1}{N},$$

we have

$$n + 1 \le \left(\frac{N+1}{N}\right)^{n-N+1} N = \left(\frac{N+1}{N}\right)^{n}\left(\frac{N+1}{N}\right)^{1-N} N$$

so

$$b_n x^n \le \left\{\left(\frac{N+1}{N}\right)^{1-N} NBq/x\right\} q^n \left(\frac{N+1}{N}\right)^n = Cr^n,$$

where C is the expression in braces. Thus, $\tilde{S}(x)$ is eventually geometrically dominated for x. ∎

Corollary 2.14. *If $S(x)$ is peaceful, then $\tilde{S}(x)$ is continuous for all $x \ge 0$.*

Proof: If $S(x)$ is peaceful, then $\tilde{S}(x)$ is peaceful by Lemma 2.13, and then $\tilde{S}(x)$ is continuous by Lemma 2.8. ∎

Theorem 2.15. *If $S(x)$ is peaceful, then $S(x)$ is differentiable for all $x \ge 0$, and $S'(x) = \tilde{S}(x)$.*

Proof: We have to show that for every $x_0 \geq 0$ and every $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\left| \frac{S(x) - S(x_0)}{x - x_0} - \tilde{S}(x_0) \right| < \varepsilon$$

whenever $0 < |x - x_0| < \delta$ and $x \geq 0$. Since $\tilde{S}(x)$ is continuous at x_0, there is a $\delta > 0$ so that $|\tilde{S}(x) - \tilde{S}(x_0)| < \varepsilon$ whenever $|x - x_0| < \delta$ and $x \geq 0$. Now

$$\frac{S(x) - S(x_0)}{x - x_0} = \frac{1}{x - x_0} \left(\sum_{n=0}^{\infty} a_n x^n - \sum_{n=0}^{\infty} a_n x_0^n \right)$$

$$= \frac{1}{x - x_0} \sum_{n=1}^{\infty} a_n (x^n - x_0^n) = \sum_{n=0}^{\infty} a_{n+1} \frac{x^{n+1} - x_0^{n+1}}{x - x_0}.$$

We have the identity $x^{n+1} - x_0^{n+1} = (x - x_0)(x^n + x^{n-1}x_0 + x^{n-2}x_0^2 + \cdots + x_0^n)$, so $(x^{n+1} - x_0^{n+1})/(x - x_0) = x^n + \cdots + x_0^n$. (There are $n+1$ terms.)

Suppose that $x_0 < x$. Then we see that, setting $y_n = (x^{n+1} - x_0^{n+1})/(x - x_0)$,

$$(n+1)x_0^n \leq y_n \leq (n+1)x^n,$$

and so

$$\sum_{n=0}^{\infty} (n+1)a_{n+1}x_0^n \leq \sum_{n=0}^{\infty} a_{n+1} y_n \leq \sum_{n=0}^{\infty} (n+1)a_{n+1}x^n,$$

i.e.,

$$\tilde{S}(x_0) \leq \frac{S(x) - S(x_0)}{x - x_0} \leq \tilde{S}(x),$$

and so

$$0 \leq \frac{S(x) - S(x_0)}{x - x_0} - \tilde{S}(x_0) \leq \tilde{S}(x) - \tilde{S}(x_0) < \varepsilon$$

(where the last inequality is true by our choice of δ). A similar argument holds if $x < x_0$. ■

We now arrive at our desired conclusion:

Corollary 2.16. *Let $E(x) = 1 + x + (x^2/2!) + (x^3/3!) + \cdots$. Then $E(0) = 1$ and $E'(x) = E(x)$ for all $x \geq 0$, i.e., $E(x)$ is a function satisfying the basic existence theorem (BET).*

Proof: We see $E(0) = 1$ by direct substitution. By Theorem 2.15, $E(x)$ is differentiable for all $x \geq 0$ and $E'(x) = \tilde{E}(x) = E(x)$. ■

Remark 2.17. Combining Definition 1.5 and Corollary 2.16, we see that

$$\exp(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \qquad \text{for } x \geq 0$$

and $\exp(x) = 1/\exp(-x)$ for $x \leq 0$. This gives us a method of calculating $\exp(x)$. For example, taking the sum of the first twelve terms of this series for $x = 1$ gives the approximation 2.71828182 for e.

Remark 2.18. We may generalize the results of this section as follows: Let $A > 0$ and say $S(x)$ is *A-peaceful* if it satisfies Definition 2.5 a) and Definition 2.5 b) for $0 \le x < A$. Then the results of this section all hold for A-peaceful series with the restriction that $0 \le x < A$ in their conclusions. In Lemma 2.6, the hypothesis becomes that $a_n \ge 0$ for all n, $a_n > 0$ for n sufficiently large, and $\lim_{n \to \infty} a_{n+1}/a_n = 1/A$, and the conclusion becomes that $S(x)$ is A-peaceful. The proofs are almost the same.

In particular, consider the series

$$L(x) = x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots.$$

It is easy to check that $L(x)$ is 1-peaceful. Then $\tilde{L}(x) = 1 + x + x^2 + x^3 + \cdots$, and we recognize this as a geometric series with sum $1/(1-x)$ for $0 \le x < 1$, i.e., $\tilde{L}(x) = 1/(1-x)$ for $0 \le x < 1$. Thus we see that $L'(x) = \tilde{L}(x) = 1/(1-x)$ for $0 \le x < 1$. On the other hand, if $f(x) = -\ln(1-x)$, then, by Proposition 1.13 and the chain rule, $f'(x) = 1/(1-x)$ for $0 \le x < 1$. We have two functions with the same derivative, so they must differ by a constant (a consequence of the mean value theorem). But $L(0) = 0 = f(0)$, so we conclude $L(x) = f(x) = -\ln(1-x)$ for $0 \le x < 1$.

This gives us a method of computing natural logarithms: For $0 < y < 1$, set $x = 1 - y$ (so $y = 1 - x$) and then $\ln(y) = -L(x) = -L(1-y)$. We know $\ln(1) = 0$. For $y > 1$, set $x = (y-1)/y$ (so $1 - x = 1/y$) and then $\ln(y) = -\ln(1/y) = L(x) = L((y-1)/y)$.

Department of Mathematics
Louisiana State University
Baton Rouge, LA 70803-4918
weintr@math.lsu.edu

In considering the epsilon-based concepts of limits, continuity, and convergence, it should be remarked that the Greek letter ε used by modern mathematicians—a notation that Cauchy originated and applied in several of his proofs—probably comes from the correspondence between '*epsilon*' and the initial letter of the French word *erreur*. In later work on probability theory, Cauchy in fact used the letter epsilon to stand for *error*. The epsilon in a modern proof may be regarded as an inheritance from the days when inequalities belonged in approximations. The epsilon notation is a reminder that, paradoxically, the development of approximations and estimates of error brought forth many of the techniques necessary for the first exact and rigorous proofs about the concepts of the calculus. Eighteenth-century mathematicians were never more exact than when they were being approximate.

Judith V. Grabiner, *The Origins of Cauchy's Rigorous Calculus*,
The MIT Press, Cambridge, 1981, p. 76.
Contributed by John P. Robertson, Anistics/Aon Worldwide Resources, NY

The Logical Structure of Computer-Aided Mathematical Reasoning

Keith Devlin

1. NEW MATH. Over the past decade or so, the professional mathematician has changed from being a person who sits at a desk working with a paper and pencil to a person who spends a lot of time sitting in front of a computer terminal. The paper and pencil are still there, but a lot of the mathematician's activities now involve use of the computer. In particular, powerful computer packages like *Mathematica* and *Maple* advertise themselves as systems for "doing math on a computer." This rapid transformation of mode of working has changed the nature of *doing* mathematics in a fundamental way. Mathematics done with the aid of a computer is qualitatively different from mathematics done with paper and pencil alone. The computer does not simply 'assist' the mathematician in doing business as usual; rather, it changes the nature of what is done. In particular, the logical structure of mathematical reasoning carried out with the aid of an interactive computer system is different from the structure of the more traditional form of mathematical reasoning. This paper, in part survey, in part a presentation of some new ideas (for none of which the author claims priority or unique observation), compares the old and the new from a logician's standpoint.

Strictly speaking, many of the points raised about computer-aided mathematics are not restricted to mathematics done on a computer. Most of the points apply, to some extent, to mathematics done in the fashion familiar to the ancient Greeks. However, the introduction of computer-aided techniques in mathematics has made those points far more salient, indeed unavoidable.

2. PROOFS AND REFUTATIONS. Ever since the ancient Greek mathematician Thales introduced the notions of theorem and proof in the sixth century B.C., *proofs* have played an important role in mathematics. It is by means of logically rigorous 'proofs' that the truths of mathematics are ultimately determined. One of the crowning glories of late nineteenth and early twentieth century logic was the formulation of a formal definition of the notion of a mathematical proof.

According to mathematical logic, a *proof* of a mathematical assertion Φ consists of a finite sequence $\phi_1, \phi_2, \ldots, \phi_n$ such that $\phi_n = \Phi$, and for each $i = 1, \ldots, n-1$, ϕ_{i+1} is either an axiom or else follows from $\phi_1, \ldots, \phi_i$ by an allowable rule of logical inference (such as modus ponens). In order to make this completely precise, the statements $\Phi, \phi_1, \ldots, \phi_n$ should be written out in a formally specified language, such as predicate logic.

However, a glance at any mathematical textbook, research monograph, or research paper chosen at random—even one in mathematical logic—will indicate that mathematicians almost never write proofs in the strict form specified by the logician's definition. In fact, the only arguments that are ever written out in strict logical fashion are proofs of extremely simple assertions given as illustrations of formal proofs, not proofs of statements of any genuine mathematical interest. There is good reason for this: for all but the simplest of mathematical assertions, a

proof written out according to the rules of formal logic would be practically impossible to follow or to understand.

The logician's notion of a formal proof is not a 'template' for the formulation of actual proofs; rather it is an 'idealization', or an 'abstract model', of a proof. A correct proof can *in principle* be formulated according to the strict rules of logic (and an invalid 'proof' cannot), and in this *asymptotic sense* the logician's notion of a formal proof relates to actual proofs in mathematics. But is there a tighter connection than this? *How closely* do the logician's proofs correspond to actual mathematical practice? Does the formal definition of a proof tell us very much about everyday mathematical arguments?

This question cannot be answered without first making precise just what is meant by the phrase 'everyday mathematical arguments'.

Ordinary mathematical practice makes at least three different, though overlapping, uses of methods of reasoning referred to as 'logical arguments'.

First of all, there are what might be called *proofs for the record*. These are the proofs that mathematicians give in their published research papers.

Second, there are the *pedagogic arguments*, the arguments mathematician use in order to explain and to convince their colleagues or their students of the truth of a particular assertion.

Third, there are the *action arguments*. These are the arguments mathematicians use in the course of solving a problem.

What are the differences between these three notions of a mathematical argument, and how does each relate to the logician's notion of a proof?

The logician's formal notion of a proof really corresponds only to proofs for the record. Arguments of the second and third kinds are generally very different. Moreover, the distinction between proofs for the record and the other kinds of argument is growing more significant as a result of the growing use of computers (and other technological tools) in mathematics.

In fact, the pedagogic and action arguments are the ones that most typify, and play the most significant role in, actual mathematical practice. Consequently, and contrary to the view implicit in classical logic, mathematics is, I suggest, far more like (Popper's description of) natural science than it resembles formal logic.

In his enormously influential book *Conjectures and Refutations: The Growth of Scientific Knowledge*, published in 1963 [**8**], the philosopher Karl Popper put forward the thesis that scientific theories are not, and cannot be, 'proved'. Rather a particular theory is at best *accepted* for the time, based on its agreement with, and prediction of, relevant observed facts, and remains accepted until refuted by some means, either a convincing counterargument or the observation of evidence to the contrary. Though nowadays widely accepted, Popper's suggestion runs counter to the generally accepted view of scientific knowledge that prevailed over the preceding centuries, whereby the primary task of the scientist was to obtain evidence to *confirm* a particular theory—the more confirming evidence that was obtained, the more secure was the theory taken to be. According to Popper, the scientist tests a theory not by confirmation but by seeking refutation. The longer a theory withstands attempts to find a refutation, the more one might feel confident in the theory and follow its predictions.

It seems clear to me that, if you examine what mathematicians actually do in their daily business—and ignore what it is they often say they are doing—then what actually constitutes a *proof* is far closer to Popper's idea than to the logician's concept of a formal proof. In the everyday world of mathematics, a *proof* is, I

suggest, nothing more nor less than an argument that:

(i) has been declared correct by a number of mathematicians acknowledged by the mathematical community to be capable of making such judgments;
(ii) has not (yet) been refuted.

Of course, the comparison between mathematics and natural science can be taken only so far. Part of Popper's thesis is that a scientific theory can *only* be an 'approximation' to reality; whether or not there is such a thing as 'truth' in science, our *knowledge* of the world can never be absolute. In contrast, mathematical truth is determinate and eternal, and our knowledge of a mathematical fact is, at least in principle, absolute. It is *in its daily practice* that mathematics resembles (I claim) Popperian science. What makes the comparison of interest is that the daily practice of mathematics is what mathematicians actually spend most of their time doing!

Incidentally, the title of this section is based on the title of Popper's book. Popper's work also inspired philosopher Imre Lakatos to name his 1976 book *Proofs and Refutations*. The parallel with the work I present in this paper is far closer to Popper's arguments about science than to Lakatos's discussion of mathematical proofs.

3. REAL DEDUCTION. Mathematical logic deals almost exclusively with a highly idealized kind of 'reasoning' that is not the reasoning that mathematicians actually employ. This is not at all a criticism of mathematical logic. On the contrary, mathematical logic succeeds brilliantly in its stated aim—to analyze the notion of mathematical proof as an idealized concept. Logic does not set out to investigate the daily praxis of mathematics. But what then of that daily praxis, which classical logic ignores? That daily praxis is the topic of this paper. What is the logical structure of the actual mathematical reasoning processes that mathematicians use in their daily work?

Specific objections to classical logic as a model of actual mathematical reasoning are that it takes no account of the following two features of actual mathematical reasoning:

- The use of diagrams and visual reasoning processes.
- Interactivity and dynamic representations.

The use of diagrams and visual reasoning have always played a major role in both pedagogic arguments and action arguments. And, as a result of the development of computers and other technological aids to reasoning, we are surely going to see a continuation of the steady rise in the use of interactivity and dynamic representation both in actual reasoning and in the subsequent presentation of arguments and results.

Of course, since the time of the ancient Greeks, mathematicians have made extensive use of diagrams, even in proofs for the record. But such uses have always been essentially as *peripherals*, to simplify the notation, to help make adequate reference to various mathematical entities, or to generally aid the reader's comprehension. It was regarded almost universally as forbidden for a proof for the record to make essential use of a diagram.

An obvious exception to the preceding statement, it might be suggested, were the ruler and compass constructions of the ancient Greeks, where the diagram was the very focus of the argument. However, even in that case one could argue that the main issue was what figures could be constructed in principle, and that the actual 'proof' did not depend on any particular diagram. Rather, the reader was

invited to convince him or herself, by abstract, internal reflection, that a certain sequence of steps would always yield the requisite figure. Indeed, it is a crucial aspect of writing up a ruler and compass 'proof', or indeed any proof in geometry, that any diagram drawn is 'generic', having no special features other than those specified in the problem. For instance, results about general triangles should not be obtained using diagrams of right-angled or equilateral triangles, a result about ellipses should not be illustrated with a diagram of a circle, and so forth.

The reason for avoiding diagrams with significant features not specified in the problem is both obvious and sensible. A valid proof cannot make use of assumptions other than those that are given (either explicitly or implicitly) in the problem. For example, a proof of a general result about triangles that depends upon the isosceles nature of the particular triangle drawn in the diagram will not be valid—at best it will be a proof that applies only to isosceles triangles. And so forth.

The danger in using diagrams—any diagram—is that there is always more information present in the diagram than stipulated in the problem. Even if, in attempting to prove a result about arbitrary triangles, a mathematician tries to draw a 'generic' triangle, it is impossible to avoid the diagram incorporating unwarranted assumptions. For example, which of the two triangles shown in Figure 1 is the more 'generic'? Each has properties not shared by the other, and is thus not typical of all triangles.

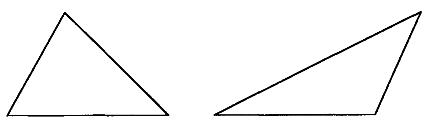


Figure 1. Which is the 'typical' triangle?

The point is, diagrams provide extremely rich and powerful representations of information. This power certainly does have the potential to mislead. But it is also the reason why mathematicians find the use of diagrams almost indispensable in their daily work. The use of the human mind to process written language, in particular written mathematical proofs, is a very recent evolutionary development, whereas humans have been using their brains to process visual information since our species first evolved. The very survival of our species depended on the human visual system, and by far the greatest proportion of the human brain is devoted to handling the input from the eyes. We are then far better equipped to handle scenes and illustrations than we are to deal with written language.

As a simple but dramatic illustration of the power of visual representations, consider Figure 2, which shows the layout of a number of blocks on a table. The task you are faced with is to determine the least number of blocks that must be removed in order to remove block L. It takes just a moment to solve the problem, based on a quick glance at the diagram. Now consider the same problem, but with the necessary information about the arrangement of the blocks given in predicate logic form:

$$\begin{aligned}&\mathrm{On}(A,B) \wedge \mathrm{On}(B,C) \wedge \mathrm{On}(C,G) \wedge \mathrm{On}(G,M) \wedge \mathrm{On}(M,E) \wedge \mathrm{On}(M,I) \wedge\\ &\mathrm{On}(G,N) \wedge \mathrm{On}(N,I) \wedge \mathrm{On}(N,Q) \wedge \mathrm{On}(C,P) \wedge \mathrm{On}(P,S) \wedge \mathrm{On}(S,Q) \wedge\\ &\mathrm{On}(S,K) \wedge \mathrm{On}(A,J) \wedge \mathrm{On}(J,F) \wedge \mathrm{On}(L,R) \wedge \mathrm{On}(R,K) \wedge \mathrm{On}(F,L) \wedge\\ &\mathrm{On}(R,O) \wedge \mathrm{On}(L,H) \wedge \mathrm{On}(H,O) \wedge \mathrm{On}(H,D) \wedge \mathrm{On}(B,F) \wedge \mathrm{On}(F,P) \wedge\\ &\forall x \forall y \forall z[\mathrm{On}(x,y) \wedge \mathrm{On}(y,z) \rightarrow \mathrm{On}(x,z)]\end{aligned}$$

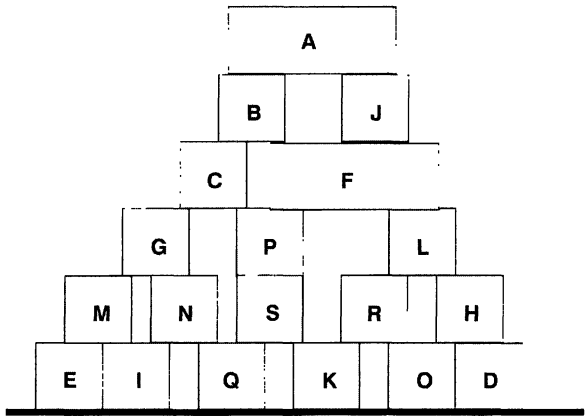


Figure 2. What is the least number of blocks that must be removed to uncover block L?

Which blocks must be removed in order to remove block L? Now the problem is much more difficult. And yet, in a very real sense, stating the problem in this formulaic fashion makes it informationally 'simpler', by providing just the pieces of information directly relevant to the problem, the information about which blocks rests on top of which other(s), together with the crucial transitivity condition on which the solution depends. Additional information given in the diagram, information not relevant to the solution, such as the lateral arrangement of the blocks (which ones are to the left/right of which others) is omitted. In fact, the diagram is cluttered with additional information irrelevant to the solution of the problem, whereas the formulaic representation provides just the information on which the solution depends. And yet, most people would say without hesitation that the problem is far simpler when presented as a diagram than when stated linguistically, using formulas. This simplicity is not informational simplicity, in the sense of the absence of spurious additional information; rather the diagramatic representation makes the problem simpler to the human mind because such a representation is ideally suited to visual processing, a task that the human visual-cognitive system performs with considerable ease.

It is interesting to note that the situation is completely reversed for a computer. It is easy to program a computer to solve the problem when it is input in the formulaic form, but writing a program that allows a computer to process visual input (say, from a video camera) is a significant challenge that computer science has by no means fully solved.

The use of diagrams may not only make the solution to a problem easier, it may also make it better. One way one proof can be 'better' than another is by providing greater insight into the problem. After all, the ultimate aim of a mathematical analysis is not 'truth' but understanding; truth is just one (very important) part of understanding. For example, consider the well known result that the sum of the first n odd integers is equal to n^2. This can be expressed by means of the identity

$$1 + 3 + 5 + \cdots + (2n - 1) = n^2.$$

A straightforward induction proof verifies this identity for all n. But where is the insight? In what way does the proof increase our understanding of what is going on?

For one thing, the very statement of the result as an algebraic identity is highly unnatural, failing to give explicit representation to all the features that make the

result interesting: that what are being summed are the first n odd integers, and that the result is the square of the number of odd integers being summed. The algebraic identity looks like just another curious, 'accidental' result of algebra; the expression in words might be a little more cumbersome, but at least it makes it clear that the result has definite interest, and expresses a 'fundamental' property of the integers.

Turning to the proof, the standard induction proof provides little by way of insight, and certainly no clue as to how the identity might have been discovered in the first place. It simply confirms the truth of the identity. Now look at Figure 3. This diagram proves the result without any need for additional explanation. The eye at once sees the crucial pattern that yields the result. The additional complexity that arises from the geometric aspects of the diagram does not lead to a more difficult proof; on the contrary, the proof becomes transparent. And far more informational.

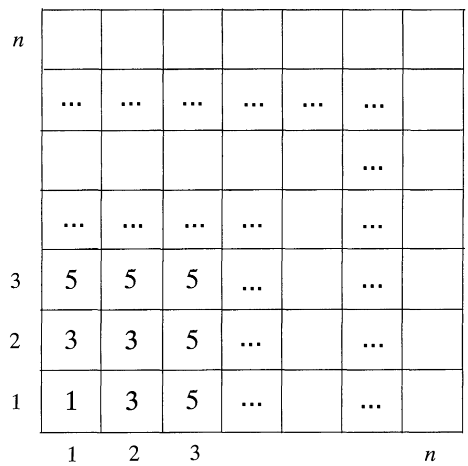


Figure 3. A proof without words or formulas.

From the perspective of the broad goal of understanding, rather than the far narrower one of establishing truth, there is a clear advantage of a transparent, explanatory proof such as the preceding one, over a technically complex but less illuminating argument. Of course, in many cases, there is little choice. Given that establishing mathematical truth is a significant part of mathematics, any proof is better than none.

However, pedagogic advice is not the main focus of this article. Rather, I am concerned here not so much with what we mathematicians *should* do but what we *actually* do. And what we do when we set about solving a problem is primarily a process of acquiring *information.* The establishment of truth may be the principal goal, but it is achieved by acquiring information about the problem. The role of truth in the construction of proofs is to motivate and guide the accumulation of the appropriate information. By concentrating on truth, indeed, by being defined in a truth-theoretic fashion, classical logic does not capture the process of (actual) proofs. Rather, what classical logic does, and does extremely well, is capture the *outcomes* of the proof process.

This distinction between truth and information becomes the more marked when mathematical analyses and arguments are carried out with the aid of a computer. Computers do not establish truth; they do not deal with truth. What computers do is store and process information. More precisely, computers store and manipulate *representations* of information—this distinction plays an important role in the analysis of reasoning I want to discuss.

4. INFORMATION AND REPRESENTATION. In order to analyze the structure of mathematical arguments viewed as processes of accumulating information, we should first ensure we have a sufficiently clear understanding of what is meant by the word 'information'. And here at once we have a problem. For, whereas we all probably have a reasonable intuitive conception of information, there is no mathematically precise definition, or at least no generally accepted definition. What there is—and what we will have to content ourselves with, at least initially—are a number of refined intuitions that can be captured in a moderately rigorous framework.

The first intuition is that information (whatever it is) is not the same as the thing in the world that represents it. What information a person, animal, computer, or other kind of agent can acquire from some object or part of the world depends on a number of factors.

For example, suppose you come across a tree stump in the forest. What information can you pick up from your find? Well, if you are aware of the relationship between the number of rings in a tree trunk and the age of the tree, the stump can provide you with the age of the tree when it was felled. To someone able to recognize various kinds of bark, the stump can provide the information as to what type of tree it was, its probable height, shape, leaf pattern, and so on. To someone else it could yield information about the weather the night before, or the kinds of insects or animals that live in the vicinity; and so on. It is not hard to think of further pieces of information that can be obtained from the stump.

Likewise, your acquisition of information stored in books depends on your knowing the language in which the book is written. You probably need more. To profit from reading an advanced book on manifolds, for example, you need to start out with a considerable knowledge of the field.

In general, information can be obtained from objects, scenes, events, environments, various kinds of physical configuration (including books and other written materials), mathematical structures, and so forth. The single term *situation* may be used to refer to any of these information bearing entities, a word whose normal meaning accords very well with some of the entities that provide information, less well with others.

In order to study mathematical reasoning that involves different forms of representation—formulas, words, diagrams, graphs, computers, and whatever—we need a formal machinery to handle the manner in which:

(i) situation can represent information;
(ii) a situation can represent different information;
(iii) different situations can represent the same information;
(iv) information can be represented in a distributed fashion, involving different representations in different situations;
(v) information can be processed by manipulating representations.

These are general issues that arise in any kind of information scenario. The only special feature in the case of mathematical reasoning is that the situations of

primary concern are mathematical structures and the various representational media involved.

Notice that any framework that can handle these five issues must surely include a notion of information that is independent of representation. Thus, what we need is a formal apparatus to handle situations and suitable notions of 'information' and 'representation'. An appropriate theory was developed during the 1980s, starting with some ideas of Jon Barwise and John Perry, who called the new framework *situation theory*. The brief summary of this theory given in the following section is designed to provide a general sense of the theory at an intuitive level. The reader who wants to gain a usable understanding of the theory should consult **[6]** for a far more extensive coverage.

5. SITUATION THEORY. *Situation theory* is a mathematical theory designed to provide a framework for the study of information. The theory takes its name from the mathematical device introduced in order to take account of context and partiality. A *situation* can be thought of as a limited part of reality. Such parts may have spatio-temporal extent, or they may be more 'abstract,' such as fictional worlds, contexts of utterance, problem domains, mathematical structures, databases, or Unix directories.

In situation theory, information is always taken to be information *about* some situation, and is taken to be in the form of discrete items known as *infons*. These are of the form

$$\langle\langle R, a_1, \dots, a_n, 1\rangle\rangle, \langle\langle R, a_1, \dots, a_n, 0\rangle\rangle$$

where R is an n-place relation and $a_1, \dots, a_n$ are objects appropriate for R (often including spatial and/or temporal locations). These may be thought of as the informational item that objects $a_1, \dots, a_n$ do, respectively, do not, stand in the relation R. The final constituent in an infon, 0 or 1, is called its *polarity*.

Infons are 'items of information'. They are not things that in themselves are true or false. Rather a particular item of information may be true or false about a situation. Though I have here introduced them as composite objects, built from relations and other objects, infons may also be defined as equivalence classes of representations under the equivalence relation of 'informational equivalence', where a representation is defined as an ordered pair consisting of a situation and a constraint. This idea is explained more fully in **[6]**.

Infons may be combined in various ways to produce more complex informational items known as *compound infons*, described momentarily.

Given a situation, s, and an infon (or compound infon) σ, we write

$$s \models \sigma$$

to indicate that the infon σ is 'made factual by' the situation s, or, to put it another way, that σ is an item of information that is true of s. The official name for this relation is that *s supports σ*. The facticity claim $s \models \sigma$ is referred to as a *proposition*.

If s is a situation and Σ is a finite set of infons, we write

$$s \models \Sigma$$

to mean that $s \models \sigma$ for all infons $\sigma \in \Sigma$.

There are several ways that infons may be combined to form *compound infons*. We need two particularly simple ways for this account.

First of all, we can form the *conjunction*, $\sigma \wedge \tau$, of two infons (or of two compound infons), σ, τ. The informational meaning of the conjunction operation

is fairly clear: the compound infon $\sigma \wedge \tau$ is the informational 'item' that comprises both σ and τ.

By definition, for any situation, s, we have

$$s \models \sigma \wedge \tau \quad \text{if and only if} \quad s \models \sigma \quad \text{and} \quad s \models \tau.$$

Likewise, the *disjunction* of two infons (or compound infons), σ, τ, is a compound infon, $\sigma \vee \tau$, such that for any situation, s,

$$s \models \sigma \vee \tau \quad \text{if and only if} \quad s \models \sigma \quad \text{or} \quad s \models \tau \quad \text{(or both)}.$$

The term *constraint* refers to the abstract mechanism by which a situation can encode, represent, and yield information and a person or agent can extract that information. Constraints may be natural laws, conventions, analytic rules, linguistic rules, empirical, law-like correspondences, or whatever.

Constraints operate at what situation theorists call the 'type level'. I can illustrate types and constraints by means of a simple example.

Suppose you are driving a car in an unfamiliar city and you come to an intersection controlled by traffic lights. You have never before been at that particular intersection, and have never before seen those particular traffic lights. Nevertheless, you know how to behave: if the lights are red, you stop, if they are green, you proceed. Your behavior cannot be based purely on the circumstances of the particular situation you find yourself in. That is a brand new situation you have never encountered before. Rather the situation is of a type you are very familiar with. You know how to behave in *any* intersection-controlled-by-traffic-lights situation. Your behavior is determined by the situation *type*. You are familiar with a *constraint* that links one type of situation, the type where the light is at red, with another type of situation, the type where you stop the car. In this case, the constraint is codified into the traffic laws, but many constraints are not.

For an example of a constraint that is not part of our system of laws, suppose you look up and see dark clouds in the sky. You say to yourself, "It looks like it might rain today." On the basis of one situation, the cloudy sky right now, you infer information about another, future situation, the weather in that region later in the day. The basis for this inference is that you are aware of a systematic relation (a constraint) between skies of a certain type and subsequent weather of a certain type. When the sky is of the type 'darkclouds', it is often the case that the weather is subsequently of the type 'raining'. The actual dark cloud formation that you see on that particular occasion does not, in itself, tell you anything. It is a one-off event. It is by virtue of the sky being of a recognizable type that you can obtain the information you do. If we were not able to recognize types, the world would always be presented to us anew, and we would be unable to make any reliable inferences based on prior knowledge or past experiences.

The ability to recognize types of things lies at the basis of much of human cognition and communication. Humans are type recognizers. So too, it appears, are various animals—bees seem able to recognize the types of certain flowers, and pet cats and dogs seem able to recognize the type of feeding bowls and the type of doorways.

Many of the words in our language refer to types: types of object, types of action, etc. For example, nouns that denote things do so by making reference to types: 'house' refers to any house, not one particular house; 'car' refers to any car, not one particular car; 'mountain' refers to any mountain, regardless of its exact shape or size or location; and so on. Similarly for verbs: 'walk' refers to any walking action by any person or legged animal; 'run' refers to any running action;

'climb' refers to any climbing action; etc. Such nouns and verbs can be used on any particular occasion to refer to a particular thing or action, but such reference is only possible because their 'meaning' refers to *types* of things or actions.

Situation theorists reify such common features of situations as *situation types*. The 'type' of all traffic-light-at-red situations is what all those situations have in common—whatever that is. This sounds a bit mysterious, and it is. But it is no more mysterious than the counting numbers that we learn to use when we are two or three years old, 1, 2, 3, etc. When you ask what the number 5 really 'is', the only answer is that it is what all collections of five objects have in common, five apples, five oranges, five elephants, five languages, etc.

Likewise, situation theorists define constraints in a formal fashion as relations between pairs of situation types.

For a mathematical example, groups constitute a type. The integers under addition and the nonzero rationals under multiplication are two situations of the type 'group'. Whenever you learn that a particular mathematical structure (i.e., a particular situation) is of the type 'group', you immediately know a lot about the properties of that situation—even if it is a situation (i.e., a structure) you have never encountered before. (Just as for the traffic lights example.)

The eventually successful search for a classification of the finite simple groups can be regarded as an attempt to establish a constraint to the effect that any structure (situation) of the type 'finite simple group' must be of a type among a certain collection of other types (type of 'cyclic group', type of 'alternating group', etc.)

The machinery of situations, types, and constraints allows us to analyze information. Given an appropriate constraint, anything can be used to store information. A knowledge or an awareness of the relevant constraint, or an adaptation to it, is what enables a person to acquire the information represented by way of that constraint. Constraints also facilitate reasoning and communication, including, of course, reasoning and communication in mathematics. *Inference* and *deduction* are activities whereby certain facts (items of information) about some situation are used to extract additional information that is in some sense implicit in those initial facts. For example, familiarity with the constraint that smoke comes from fire enables a person or an animal to infer the presence of fire from the appearance of smoke.

From the perspective of situation theory, 'logic' is the study of certain kinds of constraints, with classical, first-order, predicate logic being the study of those particular constraints established by the axioms of first-order logic.

Notice that the situation-theoretic conception of deduction makes no mention of language. This is quite different from the case of classical logic, in which language (i.e., first-order predicate language) plays a pre-eminent role. Language remains an important means of performing inference, but it is by no means all-embracing. People make non-linguistic inferences every day. As you prepare to leave for work in the morning you notice the dark clouds in the sky and reach for your raincoat, having (correctly) inferred that (a) there is a strong possibility of rain, and (b) if you do not take your raincoat you may get wet. No use of language is involved in making these inferences (except subsequently, perhaps, should you have cause to reflect upon why you acted as you did). Indeed, similar kinds of inference are made routinely all the time by animals and organisms that do not have any linguistic capacities.

Situation theory provides a rich mechanism to generate types and construct typed parameters. For the present purposes, we need just one type-formation

operation. Given a situation parameter $\dot{s}$ (dots generally denote parameters) and a compound infon σ, there is a corresponding *situation-type*

$$[\dot{s}|\dot{s} \models \sigma]$$

the *type* of situation in which σ obtains. For example,

$$[\dot{s}|\dot{s} \models \langle\langle \text{run}, \text{Max}, 1\rangle\rangle]$$

denotes the type of situation in which Max is running.

As we have observed, constraints are binary relations that link two types. The class of constraints that we need for our present analysis of mathematical reasoning are all of the form

$$S \Rightarrow S'$$

where S and S' are situation types. This constraint says that whenever there is a situation s of type S, there is a related situation s' of type S'. For many constraints, s' is either the same as s or else an extension of s. This is particularly true of many mathematical theorems, which are of the form: if s is a structure (i.e., a situation) of type S, then s is a structure of type S'.

6. THE BLOCKS PROBLEM. As an illustration of how situation theory may be used to analyze the solution to a mathematical problem, consider the stacked blocks problem introduced in Section 3. Let s be the situation represented by the diagram in Figure 2. It does not matter in this case whether s is an actual stack of blocks or an imaginary one; in fact, one can even let s be the diagram itself. The point is, s is extremely rich in information, most of which is logically—but perhaps not cognitively—irrelevant to the problem of determining how many blocks must be removed in order to uncover block L.

The solution to the problem depends (logically) upon the following two propositions:

1. $s \models \langle\langle \text{On}, A, B, 1\rangle\rangle \wedge \langle\langle \text{On } A, J, 1\rangle\rangle \wedge \langle\langle \text{On}, B, F, 1\rangle\rangle \wedge$
   $\langle\langle \text{On}, J, F, 1\rangle\rangle \wedge \langle\langle \text{On}, F, L, 1\rangle\rangle$
2. $\mathscr{M} \models [\dot{s}|\dot{s} \models \langle\langle \text{On}, \dot{x}, \dot{y}, 1\rangle\rangle \wedge \langle\langle \text{On}, \dot{y}, \dot{z}, 1\rangle\rangle] \Rightarrow [\dot{s}|\dot{s} \models \langle\langle \text{On}, \dot{x}, \dot{z}, 1\rangle\rangle]$

The first proposition provides the minimal stacking information required to solve the problem. In the second, $\mathscr{M}$ denotes the framework of elementary mathematical reasoning. Notice that the situation $\mathscr{M}$ has not been precisely defined—at least not by listing all of the rules mathematicians use in solving problems. In terms of describing the logical solution to the blocks problem, there is no need to do so. Proposition 2 supplies the crucial rule of logic required to solve this particular problem.

It can of course be argued that the solution requires additional rules. But such an observation applies to *any* argument or proof. There is, quite literally, no end to the degree of refinement and 'further detail' that can be asked for. Remember, in their mammoth work *Principia Mathematica*, it took Whitehead and Russell 362 pages of strict, formal logical development before they were able to prove $1 + 1 = 2$. And it is still possible to ask for more detail than they supplied.

One of the features of situation theory that makes it particularly suited to modeling mathematical reasoning is that its formalism can represent both the mass of general background rules and know-how that are required in order to carry out any argument (or any other cognitive task) and the particular rules and facts that are directly pertinent to the argument. The general background is captured by a situation. The pertinent rules and facts in that background are captured by infons supported by the background situation. Just which rules are regarded as pertinent

is a matter for the mathematician to decide—what passes for an adequate level of detail for the experienced, professional mathematician differs from what is required by a novice. Situation theory does not resolve the issue of how fine the analysis should be. (To some extent, classical logic does determine the level of analysis. It is because that level is far finer than most mathematicians are able to tolerate—apart perhaps from Whitehead and Russell—that mathematicians never write up their proofs in predicate logic.) Moreover, situation theory—indeed any other mathematical theory—does not have anything to say about how people come to recognize in a diagram or figure the crucial information needed to solve a particular problem; that is an issue for the cognitive psychologist to investigate. What the situation-theoretic analysis does tell you in the case of the blocks example is that the solution depends on a mixture of visual reasoning and abstract, logical deduction. The visual reasoning involves the situation s, and yields the pertinent information provided by the proposition 1. The abstract reasoning involved is based most significantly on the constraint in proposition 2.

So far, our analysis of the solution to the blocks problem has simply identified and isolated the crucial information—though it should be remembered that we are regarding mathematical arguments as consisting primarily of the acquisition of appropriate information, so the identification of the relevant information obtained at each stage of the argument is the most significant feature of our analysis. Suppose we wanted to take the analysis a stage further, and see how proposition 1 arises.

In my own case, I solve the blocks problem as follows. I first identify the target block L. I then restrict my visual attention to the subsituation s' of s that comprises all of the blocks that seem to be 'above' block L. I do not decide exactly which blocks are in the subsituation s' and which are not. At this stage, that level of precision is irrelevant to the solution. The vague, unspecified description of s' I just gave is all that I require. Notice that the same is true of the entire stack situation s. The solution does not require that you identify every single block. For instance, do you know, without looking, where block Q is located? Is there a block labeled Z? Both of these items of information would be part of a complete, extensional description of the stack, but they are not required in order to solve the problem given. The important feature of s' as far as the solution is concerned is that:

3. $s' \models \langle\langle \text{On}, A, B, 1\rangle\rangle \wedge \langle\langle \text{On}, A, J, 1\rangle\rangle \wedge \langle\langle \text{On}, B, F, 1\rangle\rangle \wedge$
$\langle\langle \text{On}, J, F, 1\rangle\rangle \wedge \langle\langle \text{On}, F, L, 1\rangle\rangle$

This is proposition 1 with s' in place of s. Proposition 3 implies proposition 1 by virtue of a fundamental fact about infons known as *persistence*. This property says that if u is a situation and σ is an infon, and if $u \models \sigma$, then for any situation u' that extends u, $u' \models \sigma$. The property of persistence has deep consequences concerning the structure of relations and the requirements that have to be satisfied by infons. See **[6]** for details.

Notice that the restriction of attention from the entire stack s to the substack s' is something the human visual-cognitive system does directly and with ease. Such aspects of visual reasoning cannot be adequately captured by traditional logical formalisms because they are extensionally imprecise. Their enormous efficiency comes in large part from the fact that they do not attempt to resolve issues that are ultimately irrelevant to the solution.

By augmenting proposition 1 with proposition 3, we obtain a more fine-grained analysis of the solution to the problem. Notice that the constraint in proposition 2

can be applied either to situation s or to situation s'. In my solution to the problem, I never try to apply the constraint to the large situation s. I do however apply it to the much smaller situation s', where there are far fewer blocks to consider.

At this stage, in my case, I am able to solve the problem by inspection. I certainly do not resort to a 'search'. Though I never bother to calculate its exact size, situation s' certainly contains no more than seven blocks, and it is a well established fact of cognitive psychology (first observed by the psychologist George Miller in 1956) that the human visual-cognitive system is able to process structural and quantitative information about any collection of seven or fewer objects. In my case, the key step in the solution to the problem is the restriction of attention to the substack s', which is small enough that I can avoid having to resort to a search procedure. Situation theory is able to model this key step in a way that is not possible with classical logic or other systems that require complete extensional precision.

Of course, it would be possible to continue to refine the analysis. But from the point of view of illustration, it should now be reasonably clear how situation theory is used in such an analysis.

7. THE STRUCTURE OF COMPUTER-AIDED PROOFS. The example considered in the previous section did not involve any use of a computer, but it illustrated the key feature of computer-aided reasoning: the use of more than one means to represent and manipulate information. When a mathematician makes use of a modern computer system for doing mathematics, such as *Mathematica* or *Maple* or perhaps some more specialized package, not only is the process very clearly one of acquiring and processing information (rather than establishing truths), but the reasoning can be very different from mathematical reasoning performed without such aids.

Of course, the use of the computer could be fairly superficial, say just to draw an initial diagram to illustrate a traditional, pencil-and-paper type proof, or to perform some arithmetical calculations. Though mathematical arguments carried out in this fashion may still be analyzed using the techniques outlined in this article, I am more interested in reasoning processes in which the computer is used in a significant way, as an integral part of the reasoning process.

For example, a fast, interactive graphics facility can enable the mathematician to draw not just one but a whole series of graphs or diagrams, to enlarge part of a graph and zoom in on regions of particular interest, to rotate or reflect a surface, to set up a short 'movie' that illustrates the effect of a particular parameter change, and so forth. These are usually done in conjunction with more traditional forms of reasoning. A mathematician skilled in the use of such systems is able to move from one medium to another in a smooth, effortless fashion, going from pencil and paper to computer screen and back to pencil and paper again. These changes in representational/reasoning medium do not change the problem; there remains a single problem, and the entire process results in an accumulation of a single body of information about that problem. What changes during the course of a proof is the medium in which the informational is represented and processed; first one, then another, then another, sometimes a mixture of two or more representations at the same time. Except for the dynamic aspect of computer representations, this is precisely the feature of reasoning with a diagram that was investigated in the previous section. Situation theory is well-equipped to handle dynamic issues, since time can be one of the arguments in an infon.

Classical logic and natural deduction have no mechanism to handle multi-representational reasoning, since at most they can capture *what* is represented, not *how* it is represented. And, of course, one enormously significant feature of any kind of diagramatic reasoning process is that the diagram contains far more information than is *logically* required for the argument, which is something else that classical logic cannot reflect but situation theory can handle with ease.

When a mathematician starts to solve a problem, the data given establish both a problem domain—perhaps a certain kind of mathematical structure such as a group or the real number system, or maybe a stack of blocks or an arrangement of rods on a table—and some initial information about that problem domain. We shall represent the problem domain as a situation, say d, and the given data by a finite set Σ of infons. The starting point thus constitutes a finite set of propositions $d \models \Sigma$ (i.e., $d \models \sigma$ for every σ in Σ). In these terms, the goal may be to demonstrate that d satisfies some particular property τ, which amounts to establishing the proposition $d \models \tau$ or the goal may be to show that there is some further situation e having a certain property, $e \models \tau$ We then model the reasoning process as the establishment of a series of sets of propositions $d_1 \models \Sigma_1$, $d_2 \models \Sigma_2, \dots,$ $d_n \models \Sigma_n$ where $d_1 = d$, $\Sigma_1 = \Sigma$, and the final proposition is the goal (so $\Sigma_n = \{\sigma_n\}$ for a single compound infon σ_n). The situation under consideration at each stage in this process is called a *case*. If the compound infons in the set Σ_i (at stage i of the deduction) are contradictory, we say the case d_i is *terminal.* Any nonterminal case that is subsumed (see presently) by one or more later cases is said to be *closed*; any other nonterminal case is said to be *open.*

In carrying out the argument, any stage $d_i \models \Sigma_i$ can be advanced in one of two ways: either by acquiring additional information about the situation d_i (i.e., adding further infons to the set Σ_i) or else by turning attention to one or more different situations d'_i, d''_i, etc., perhaps by breaking d_i into subcases.

Another way to advance the proof is to take a number of stages $d_i \models \Sigma_i$, $d_j \models \Sigma_j$, $d_k \models \Sigma_k, \dots$ and subsume them by a single new stage $d_t \models \Sigma_t$ that manages to combine all the information in the stages subsumed. This is what is done when we have exhausted all cases in a 'proof by cases' argument.

Notice that the linear order of the cases and propositions in an argument primarily reflects the fact that the reasoning takes place in time. The logical structure need not be linear; most commonly it takes the form of a cycle-free directed graph.

Barwise and Etchemendy [**1**] have suggested that any piece of mathematical reasoning involves just five distinct principles, which I list below. They used these principles to develop an educational software system called *Hyperproof* [**2**], designed to teach logical reasoning using a combination of ordinary language, mathematical formulas, and diagrams. They refer to such reasoning as *heterogeneous reasoning*. Here are Barwise and Etchemendy's five principles of heterogeneous reasoning:

Given. Accept some initial information as given. This step gives rise to the initial case.

Assume. Given some open case d, assume something extra, thereby creating an open subcase of d. A typical use of this principle is to assume the antecedent of a conditional with the goal of deducing the consequent, thereby establishing the conditional.

Subsume. Disregard some open case if it is subsumed by all other open cases. For instance, this step is performed when all the information in a case is exhausted by its subcases.

Merge. Take the information common to a number of open cases, and call it a new open case. This principle would typically apply after a case has been broken into a number of subcases and a desired result established in each subcase.

Recognize as Possible. Given some open case, recognize it as representing a genuine possibility. This form of reasoning is typically used when a counterexample is accepted as showing that some result does not follow from the given information.

Classical logic is largely restricted to a combination of assume and subsume.

It should be stressed again that the situation-theoretic approach just outlined does not attempt to describe how a person comes to make the various decisions involved in constructing an argument—which of the five steps listed above to take, which of the available representations to use at any stage, and so forth. Situation theory does not even claim to provide a model of the actual process of deriving an argument—it is hard to imagine *any* mathematical theory being able to model mental processes of that nature. What a situation-theoretic analysis does model is the *outcome* at each stage, the outcome not in terms of truth, rather the informational outcome. But notice that steps such as deciding what kind of move to make next and choosing the representation are crucial and integral parts of the solution process. By taking account of (without modeling) the *actual* human reasoning process, situation theory does, however, add a degree of mathematical precision to a description of actual mathematical reasoning. It does not, of course, produce a totally formal account on the level of mathematical logic. As such, the application of situation theory described in this paper constitutes an instance of what in **[7]** I refer to as 'soft mathematics', a use of mathematical ideas and formalisms common in the human and social sciences, and, if Casti **[4]** is right, about to become common in applied mathematics as well.

REFERENCES

1. Barwise, J. and Etchemendy, J. Information, Infons, and Inference, in Cooper, Mukai, and Perry (editors), *Situation Theory and Its Applications, Vol.*1, Stanford University: CSLI Lecture Notes 22, 1990.
2. Barwise, J. and Etchemendy, L. *Hyperproof*, CSLI Publications, 1994.
3. Barwise, J. and Perry, J. *Situations and Attitudes*, MIT Press, 1983.
4. Casti, J. Confronting Science's Logical Limits, *Scientific American*, October 1996, pp. 102–105.
5. Dalen, D. van *Logic and Structure*, Springer-Verlag, 1980.
6. Devlin, K. *Logic and Information*, Cambridge University Press, 1991.
7. Devlin, K. *Goodbye Descartes: The End of Logic and the Search for a New Cosmology of the Mind*, John Wiley, 1997.
8. Popper, K. *Conjectures and Refutations: The Growth of Scientific Knowledge*, Basic Books, 1993.

Saint Mary's College of California
Moraga, CA 94575
devlin@stmarys-ca.edu

NOTES

Edited by **Jimmie D. Lawson**

The Wallet Paradox

Kent G. Merryfield, Ngo Viet, and Saleem Watson

1. Introduction. In his book *aha! Gotcha* [1] Martin Gardner gives an intriguing "paradox" involving money and wallets. We found that an analysis of this paradox can serve as an interesting way of utilizing some key concepts in probability. The paradox as related by Gardner is as follows:

> Each of two persons places his wallet on the table. Whoever has the smallest amount of money in his wallet, wins all the money in the other wallet. Each of the players reason as follows: "I may lose what I have but I may also win more than I have. So the game is to my advantage."

Paradoxically, it seems that the game is to the advantage of both players. Of course if one player always carries a larger amount of money than the other player, then he always loses. So we must require that the game be "fair" in some sense. In his analysis of the problem Kraitchik [2] assumes that the amount of money each person carries is uniformly (discretely) distributed between 0 and 100. He then makes a chart of the distribution of money of both players and observes that the distribution is symmetric (with respect to the diagonal) and concludes that there is no advantage. This explanation is considered unsatisfactory by Gardner since it does not explain what is wrong with the reasoning of the players. Indeed, Kraitchik's chart gives a particular example where the game is not to the advantage of either player, but does not address the source of the paradox. In this article we explore the concept of a fair game and in the process we shall resolve the paradox.

2. What are the random variables? A player says "I may lose what I have but I may also win more than I have." This is a true statement for any single trial of the game. However, the inference that "the game is to my advantage" is the source of the apparent paradox, because it does not take into account the *probabilities* of winning and losing. In other words, if the game is played many times, how often does a player win? How often does he lose? And by how much? Indeed, by considering many trials of this game, the enthusiasm of the players for winning should be tempered by the observation that when one loses one typically has more money in one's wallet.

To analyze this game probabilistically we need to know what are the relevant random variables and what are their probability distributions [3]. We are interested in the probability distributions of W_A and W_B, the amount of money that player A and B will win (or lose), respectively. We say that the game is *fair* if the expected value $E(W_A) = 0$ (equivalently $E(W_B) = 0$). To understand W_A and W_B, let X and Y be the random variables representing the amount of money in the wallets of

players A and B, respectively. According to the rules of the game W_A is given by

$$W_A(X,Y) = \begin{cases} -X, & \text{if } X > Y \\ Y, & \text{if } X < Y \\ 0, & \text{if } X = Y \end{cases}$$

and $W_B(X,Y) = -W_A(X,Y)$. These expressions make it difficult to calculate $E(W_A)$ and $E(W_B)$, since they depend on both of the distributions of X and Y in a nontrivial way. This is apparently why no quick and simple way of resolving the paradox is available.

We now consider models of the game that intuitively seem fair.

3. A fair game: independent identically distributed X and Y. Perhaps the most natural model for this game is one in which the distributions of the money in each player's wallet are the same; that is, X and Y are independent, identically distributed random variables on some interval $[a,b]$ or $[a,\infty)$, $0 \le a < b < \infty$. Thus the joint distributions (X,Y) and (Y,X) have the same density function f satisfying $f(x,y) = f(y,x)$. Now, observing that $W_A(Y,X) = W_B(X,Y)$ and exploiting the symmetry in the problem we have

$$E(W_A) = \int_a^b \int_a^b W_A(x,y) f(x,y)\, dy\, dx$$

$$= \int_a^b \int_a^b W_A(y,x) f(y,x)\, dy\, dx = \int_a^b \int_a^b W_B(x,y) f(x,y)\, dy\, dx = E(W_B)$$

where we have made the change of variables $(x,y) \to (y,x)$, whose Jacobian is 1. This combined with the observation that $W_A(X,Y) = -W_B(X,Y)$ shows that $E(W_A) = 0$; and so, by our definition, this is a fair game.

As a concrete example, suppose X and Y are jointly uniformly distributed on the unit square $[0,1] \times [0,1]$. The probability that player A wins y dollars is $1-x$. In that case $y \in (x,1]$ with mean equal to $(1+x)/2$. Player A loses x dollars with probability x. Given that player A carries x dollars in his wallet, the conditional expectation of the amount of money that he will win is

$$E(W_A | X = x) = \left(\frac{1+x}{2}\right)(1-x) - x^2 = \frac{1}{2} - \frac{3}{2}x^2.$$

Thus the expected value for W_A is

$$E(W_A) = \int_0^1 E(W_A | X = x)\, dx = \int_0^1 \left(\frac{1}{2} - \frac{3}{2}x^2\right) dx = 0.$$

It is interesting to consider special cases of this formula for the conditional expectation. Since $E(W_A | X = 1) = -1$ and $E(W_A | X = 0) = 1/2$ we see that a player carrying one dollar in his wallet should expect to lose it, whereas a player carrying nothing in his wallet should expect to gain half a dollar (the mean). Interestingly, if a player is carrying half a dollar (the mean) in his wallet, then $E(W_A | X = 1/2) = 1/8$; that is, his expectation of winning is positive.

4. "The game is *not* to my advantage". It may be tempting to think that the game would be fair if we require only that the distributions X and Y have the same

mean. But this is not always the case, as we now show. Suppose that X and Y have the joint distribution shown in the following chart.

$X \setminus Y$	0	1
0	$\frac{2}{6}$	$\frac{2}{6}$
$\frac{3}{2}$	$\frac{1}{6}$	$\frac{1}{6}$

For player A, the marginal distribution is $p(0) = 4/6$ and $p(3/2) = 2/6$ and for player B, the marginal distribution is $p(0) = 3/6$ and $p(1) = 3/6$. The mean for player A is $m_A = (0 \times 4/6) + ((3/2) \times (2/6)) = 1/2$. Similarly the mean for player B is $m_B = (0 \times 3/6) + (1 \times 3/6) = 1/2$. But the expected value of player A's winnings is

$$E(W_A) = 0 \times \frac{2}{6} + 1 \times \frac{2}{6} - \frac{3}{2} \times \frac{1}{6} - \frac{3}{2} \times \frac{1}{6} = -\frac{1}{6}$$

This shows that the game is to the advantage of player B.

It turns out that even a smaller mean does not guarantee an advantage in this game. Indeed, replacing $3/2$ in the chart by any number in the interval $(1, 3/2)$ yields an example where player A has a smaller mean than that of B. However, player A is still at a disadvantage (that is, $E(W_A) < 0$).

5. Conclusion. The concept of a fair game has to do with repeated trials (and not with any single trial) of a game. So the wallet game is properly understood in the context of the probability distributions of the money in the wallets and the expected values of winning for each player. We have shown that the game is fair if reasonable assumptions are made on these probability distributions (Sections 3) whereas the game is not fair with other assumptions on these distributions (Section 4). Moreover, our analysis may be used to determine whether the game is fair for any given pair of distributions. So in the context of probability, the paradox is resolved.

Some interesting questions remain unanswered about this problem. For instance, if we suppose that the distributions of players A and B are required to have the same means, is there a strategy that player A could adopt to have a winning edge? In other words, is there a preferred distribution (or a winning strategy)?

REFERENCES

1. Martin Gardner, *Aha! Gotcha*, W. H. Freeman and Company, New York, 1981.
2. Maurice Kraitchik, *Mathematical Recreations*, 2nd edition, Dover, New York, 1953.
3. Dennis Wackerly, William Mendenhall, and Richard Sheaffer, *Mathematical Statistics with Applications*, 5th edition, Duxbury Press, New York, 1996.

Department of Mathematics
California State University, Long Beach
Long Beach, CA 90840
kmerry@csulb.edu
viet@csulb.edu
saleem@csulb.edu

The Weierstrass Approximation Theorem and Large Deviations

Henryk Gzyl and José Luis Palacios

Bernstein's proof (1912) of the Weierstrass approximation theorem, which states that the set of real polynomials over [0, 1] is dense in the space of all continuous real functions on [0, 1], is a classic application of probability theory to real analysis that finds its way into many textbooks ([**1**] and [**2**]) and journals [**3**]. All that is invoked in Bernstein's proof (at least as presented in [**1**] and [**3**]) is Chebyschev's inequality, and if the argument is applied to a function satisfying a Lipschitz condition, the rate of convergence of the Bernstein polynomials to the function can be shown to be at least of order $1/n^{1/3}$. If instead of Chebyschev's inequality we use another probabilistic tool very much in vogue nowadays, the theory of large deviations, we can prove that the rate of convergence is at least of order $\ln^{1/2} n/n^{1/2}$. All the material used here concerning large deviations is elementary and can be found in [**1**].

Let f be a real function on [0, 1] that satisfies a Lipschitz condition, i.e., there is a constant C such that for all $x, y \in [0, 1]$

$$|f(x) - f(y)| \le C|x - y|.$$

Then we have the following:

Theorem 1 (Weierstrass approximation theorem). *For f satisfying a Lipschitz condition, there is a sequence of polynomials $p_n(x)$, where the degree of $p_n(x)$ is n, and a constant K, which depends on f, such that*

$$\|p_n - f\| \le K \frac{\ln^{1/2} n}{n^{1/2}}.$$

Here $\| \; \|$ denotes the sup norm. Since the function f is Lipschitz, it is uniformly continuous and bounded by a constant, say, M so that $\|f\| \le M$. In order to prove the theorem we need the following lemma, taken almost verbatim from [**1**], (Corollary A.7) and included for completeness:

Lemma 1. *For a binomial random variable $B(n, x)$ with n independent trials and probability of success x for each of them, and $a > 0$ arbitrary, we have*

$$P(|B(n, x) - nx| > a) \le 2e^{-2a^2/n}.$$

Proof: Let $X_1, \dots, X_n$ be independent and identically distributed random variables with

$$P(X_i = 1 - x) = x,$$
$$P(X_i = -x) = 1 - x,$$

and let $X = X_1 + \cdots + X_n$. Clearly X has distribution $B(n, x) - nx$. Since X is symmetric, it is enough to prove the one-sided inequality

$$P(X > a) < e^{-2a^2/n}. \tag{1}$$

Step 1. For all reals α, β with $|\alpha| \le 1$, we have

$$\cosh(\beta) + \alpha \sinh(\beta) \le e^{\beta^2/2 + \alpha\beta}. \tag{2}$$

Proof. This is immediate if $\alpha = 1$ or $\alpha = -1$ or $|\beta| \ge 100$. If (2) were false, the function

$$f(\alpha, \beta) = \cosh(\beta) + \alpha \sinh(\beta) - e^{\beta^2/2+\alpha\beta}$$

would assume a negative global minimum in the interior of the rectangle

$$R = \{(\alpha, \beta): |\alpha| \le 1, |\beta| \le 100\}.$$

Setting partial derivatives equal to 0, we find

$$\sinh(\beta) + \alpha \cosh(\beta) = (\alpha + \beta) e^{\beta^2/2+\alpha\beta},$$

$$\sinh(\beta) = \beta e^{\beta^2/2+\alpha\beta},$$

and thus $\tanh \beta = \beta$, which implies $\beta = 0$. But $f(\alpha, 0) = 0$ for all α, a contradiction.

Step 2. For all $\theta \in [0, 1]$ and all λ,

$$\theta e^{\lambda(1-\theta)} + (1 - \theta) e^{-\lambda\theta} \le e^{\lambda^2/8}. \tag{3}$$

Proof. Setting $\theta = (1 + \alpha)/2$ and $\lambda = 2\beta$, (3) reduces to (2).

Step 3. Let, for the moment, $\lambda > 0$ be arbitrary and let $E[\cdot]$ denote "expected value." Then

$$E[e^{\lambda X_i}] = x e^{\lambda(1-x)} + (1 - x) e^{-\lambda x} \le e^{\lambda^2/8}$$

by Step 2. Thus

$$E[e^{\lambda X}] = \prod_{i=1}^{n} E[e^{\lambda X_i}] \le e^{\lambda^2 n/8}.$$

Applying Markov's inequality,

$$P(X > a) = P(e^{\lambda X} > e^{\lambda a}) \le \frac{E[e^{\lambda X}]}{e^{\lambda a}} \le e^{\lambda^2 n/8 - \lambda a}.$$

We set $\lambda = 4a/n$ to optimize the inequality: $P(X > a) \le e^{-2a^2/n}$, as claimed. ∎

Proof of the theorem. Define the Bernstein polynomials

$$p_n(x) = \sum_{i=0}^{n} \binom{n}{i} x^i (1 - x)^{n-i} f\left(\frac{i}{n}\right).$$

Then, since $\Sigma_{i=0}^{n}\binom{n}{i}x^i(1-x)^{n-i} = 1$, and because of the lemma, we have

$$\begin{aligned} |p_n(x) - f(x)| &\le \sum_{i:\,|i-nx|\le a} \binom{n}{i} x^i(1-x)^{n-i}\left| f\left(\frac{i}{n}\right) - f(x)\right| \\ &\quad + \sum_{i:\,|i-nx|>a} \binom{n}{i} x^i(1-x)^{n-i}\left(\left| f\left(\frac{i}{n}\right)\right| + |f(x)|\right) \\ &\le \frac{aC}{n} + 2MP(|B(n,x) - nx| > a) \\ &\le \frac{aC}{n} + 2Me^{-2a^2/n}. \end{aligned} \tag{4}$$

Now we optimize the parameter a in terms of n. We consider the function

$$F(a) = \frac{aC}{n} + 2Me^{-2a^2/n},$$

set

$$0 = F'(a) = \frac{C}{n} - \frac{8Ma}{n}e^{-2a^2/n},$$

and get the exponential equation

$$ae^{-2a^2/n} = \frac{C}{8M}. \tag{5}$$

While a precise solution for this exponential equation is unavailable, we are led to the asymptotic solution

$$a = \frac{1}{2}n^{1/2}\ln^{1/2} n.$$

Replacing now in (4) $a = 1/2n^{1/2}\ln^{1/2} n$, we obtain

$$\|p_n - f\| \le \left(\frac{C}{2} + \frac{2M}{\sqrt{\ln n}}\right)\frac{\ln^{1/2} n}{n^{1/2}},$$

so $K = C/2 + 2M$ works for $n \ge 3$. ∎

The classic probabilistic proof of the Weierstrass approximation theorem, when applied to a Lipschitz function, yields instead of equation (4) the expression

$$G(a) = \frac{aC}{n} + \frac{n}{a^2}, \tag{6}$$

where the first summand follows from the Lipschitz condition and the second is due to Chebyshev's inequality. Optimizing $G(a)$ (a much easier task than optimizing $F(a)$ above) yields $a \sim n^{2/3}$ and therefore, inserting $a = n^{2/3}$ into (6) (this is the choice of a in **[1]**, by the way) yields

$$\|p_n - f\| \le G(n^{2/3}) = O\left(\frac{1}{n^{1/3}}\right),$$

a weaker result, which justifies the effort of using the large deviation inequality.

How does our rate of convergence compare with the ones found in standard textbooks on approximation theory? In **[4]**, for instance, it is mentioned that if $f(x) = x^2$ and $p_n(x)$ is the Bernstein polynomial for this function, then $\|p_n - f\| = 1/4n$. In general, this rate cannot be expected for all functions, and it is an exercise in **[4]** to prove that if f is twice continuously differentiable, then the error

of the approximation satisfies the bound

$$\|p_n - f\| \le \frac{1}{8n}\|f''\|.$$

This rate is better than ours, but the assumption that f be twice continuously differentiable is much more restrictive than our Lipschitz condition.

REFERENCES

1. N. Alon and J. Spencer, *The Probabilistic Method*, Wiley, New York, 1992.
2. K. L. Chung, *A Course in Probability Theory*, 2nd ed., Academic Press, New York, 1974.
3. K. M. Levasseur, A probabilistic proof of the Weierstrass approximation theorem, this MONTHLY 91 (1984) 249–250.
4. M. J. D. Powell, *Approximation Theory and Methods*, Cambridge University Press, Cambridge, 1981.

Facultad de Ciencias
Universidad Central de Venezuela
Caracas, Venezuela

CESMa
Universidad Simón Bolívar
Apartado 89,000
Caracas, Venezuela
jopala@cesma.usb.ve

From the MONTHLY, Volume 4, 1897:

A *Brief Introduction to the Infinitesimal Calculus*. Designed Especially to Aid in Reading Mathematical Economics and Statics. By Irving Fisher, Ph.D., Assistant Professor of Political Science in Yale University, Co-author of Phillips and Fisher's Elements of Geometry. 12mo. Cloth. 84 pages. Price, 75 cents. York and London: The Macmillan Co.

This little work on the Calculus will be received with joy by a great army of students, teachers, and professors, who have lacked the time and courage to attack some of the more exhaustive works on the subject yet felt the need of a knowledge of the Calculus in order to enable them to read with intelligence the highest authorities on economic as well as other subjects. Dr. Fisher has prepared this little work with a special view of the needs of this class of students. Any one with a clear mind can very easily read and understand every sentence in this book. There is no metaphysical speculation nor obscure statements made in establishing its first principles. pp. 261–262

65. Proposed by GEORGE LILLEY, Ph.D., LL.D., Portland, Oregon.

A string is wound spirally 100 times around a cone 100 feet high and 2 feet in diameter at the base. Through what distance will a duck swim in unwinding the string keeping it taut at all times, the cone standing on its base and at right angles to the surface of the water? p. 62

66. Proposed by J. K. ELLWOOD, A.M., Principal of Colfax School, Pittsburgh, Pennsylvania

Around the top of a conical frustum—base 5 feet, top 1 foot, altitude 100 feet—is wound a rope 100 feet long and 1 inch thick. It is unwound by a hawk flying in one plane. How far does Mr. Hawk fly? p. 62

THE EVOLUTION OF...

Edited by **Abe Shenitzer**
Mathematics, York University, North York, Ontario M3J 1P3, Canada

On the Historical Development of Infinitesimal Mathematics

translated by Abe Shenitzer with the editorial assistance of Hardy Grant

Detlef Laugwitz

PART II. THE CONCEPTUAL THINKING OF CAUCHY[1]

5. CAUCHY 1821: CONTINUITY AND CONVERGENCE. Next to Fourier and Simeon Denis Poisson (1781–1840), Augustin-Louis Cauchy (1789–1857) was one of the most prolific workers in the area of the new mathematical physics. Beginning in 1816, he taught at the Ecole Polytechnique. From its establishment in 1795, this school provided a demanding basic course of study in mathematics that was to become a model for university education. There were no usable textbooks (apart from the one by Lacroix), and Cauchy had to write outlines for his lectures. Far from being learning aids for the students, they were outlines of the foundations of analysis. Like so many textbook writers after him, Cauchy thought more about scientific research than about the teaching of beginners. But this fact resulted in a fundamental reshaping of analysis. Cauchy initiated the elimination of algorithmic thinking by replacing it with conceptual thinking.

Functions were no longer given by expressions. In his Cours d'analyse of 1821, Chapter I on Real Functions opens with the following sentence: "If variable magnitudes are so interrelated that prescribing the value of one of them makes it possible to infer the values of all the others, then we ordinarily think of these magnitudes as expressed in terms of one of them, which then bears the name of an independent variable; and the magnitudes expressed in terms of the independent variable are what one calls functions of this variable." And Chapter II begins with the sentence: "One says that a variable magnitude becomes infinitely small if its absolute value decreases indefinitely, so that it converges to the limit zero."

These descriptions cannot be regarded as definitions from which one can deduce theorems. Rather, these are clarifications of usage. It is hard to tell whether or not Cauchy distances himself from the conception of a function as an "analytical expression." What counts is the way one operates, and this can be seen in the context of continuity of functions. Inferences are drawn from the conceptual property and not from the representation by means of a particular expression.

[1]In a recent MONTHLY article [**103** (1996) 846–853] Roman Kossak tries to explain "what are infinitesimals and why they cannot be seen." I hope readers of my paper realize that Cauchy knew very well what infinitesimals are and that they *can* be seen.

Continuity of a function $f(x)$, given on an interval, is defined in para. 2 of Chapter II: Let α be an infinitely small increment of x. Then $f(x)$ is said to be continuous in this interval "if the absolute value of $f(x+\alpha) - f(x)$ decreases indefinitely together with that of α. In other words, the function $f(x)$ is continuous with respect to x within the given bounds if, within these bounds, an infinitely small increment of the variable always results in an infinitely small increase of the function itself." Cauchy italicizes this alternative definition.[2]

One gets the impression that "infinitely small" is just an abbreviation for "tending to the limit 0." But one must check how Cauchy works with this notion. And so we look at his Theorem I on continuous functions: If $f(x, y)$ is continuous with respect to each of the variables x and y in the vicinity of (x_0, y_0), then $f(x, y)$ has the limit $f(x_0, y_0)$ if x and y converge to x_0 and y_0 respectively.

The statement says nothing about infinitely small magnitudes, and there are counterexamples such as $f(x, y) = xy/(x^2 + y^2)$ for $(x, y) \neq (0,0)$, and $f(0,0) = 0$. Here f is continuous with respect to y for every fixed real x and continuous with respect to x for every fixed y. But if we put $x = y = t$ and let t tend to 0, then we obtain $1/2 \neq 0$. How does Cauchy prove this obviously false theorem? Let α and β denote infinitely small magnitudes. Then $f(x+\alpha, y+\beta) - f(x, y) = [f(x+\alpha, y+\beta) - f(x+\alpha, y)] + [f(x+\alpha, y) - f(x, y)]$ is infinitely small. This is so because the expression in the first bracket is an infinitely small magnitude due to the continuity with respect to the second variable, and the expression in the second bracket is such because of the continuity with respect to the first variable.

What happens to our counterexample and the first of the two brackets for $x = y = 0$? The counterexample reduces to $\alpha\beta/(\alpha^2 + \beta^2)$ and this is not always infinitely small when α and β are; a relevant case is $\alpha = \beta \neq 0$. On other occasions, Cauchy readily gives counterexamples, and it is safe to assume that he had seen an example like the one just given. If that is so, then the interpretation to be given to his assumption is that $f(x, y)$ is continuous with respect to y for every fixed $x = x_1 + \alpha_1$, with x_1 real and α_1 a fixed (!) infinitely small magnitude. But this seems to contradict the initial description of infinitely small magnitudes as variables that converge to 0!

Cauchy's Theorem I on convergent series of continuous functions places us in a similar dilemma. Let $s_n(x)$ denote the necessarily continuous partial sums of a series of continuous functions, and suppose that this sequence converges for all x in an interval to a limit function $s(x)$. Then

$$\begin{aligned} s(x+\alpha) - s(x) = {} & [s(x+\alpha) - s_n(x+\alpha)] \\ & + [s_n(x+\alpha) - s_n(x)] + [s_n(x) - s(x)]. \end{aligned}$$

If α is an infinitely small magnitude, then, because of the continuity of s_n, the expression in the second bracket on the right is always infinitely small. Because of the convergence at x, the absolute value of the expression in the last bracket is certainly less than a given real $\epsilon > 0$ for $n \geq N(\epsilon)$. If this is also true for the expression in the first bracket, then we obtain $|s(x+\alpha) - s(x)| < 3\epsilon$ for arbitrary

[2] A reader influenced by today's school mathematics may be shocked by the expression 'the function $f(x)$,' which we use in this paper in connection with Cauchy. This expression is entirely unobjectionable if, as has usually been the case since Cauchy, we use the last letters $x, y, z, u, v, \ldots$ of the alphabet as symbols for variables, and indexed letters $x_0, y_1, \ldots$, or the first letters $a, b, c, \ldots$ of the alphabet, as symbols for fixed values. The domains of definition, in our case invariably intervals, are inferred from the context. The fear of interchanging function and functional value is a problem for beginners and not for advanced and cautious users.

positive real ϵ. Hence $s(x + \alpha) - s(x)$ is an infinitely small magnitude. But then $s(x)$ is a continuous function. Cauchy states this as a theorem: "If the different terms of a series are functions of a variable x which are continuous with respect to this variable in the vicinity of a particular value for which the series is convergent, then the sum of the series is likewise a continuous function of x in the vicinity of this particular value."

Here is the original text: "Lorsque les différens termes de la série sont des fonctions d'une même variable x, continues par rapport a cette variable dans le voisinage d'une valeur particulière pour laquelle la série est convergente, la somme s de la série est aussi, dans le voisinage de cette valeur particulière, fonction continue de x."

We see that the assumed continuity in the vicinity of $x = x_0$ is used to prove that $s_n(x_1 + \alpha_1)$ converges to $s(x_1 + \alpha_1)$ for a real fixed x_1 and an infinitely small fixed α_1.

It is certain that Cauchy knew the relevant counterexamples—Fourier series of discontinuous functions—very well, for he had used them in his research for years. More simply, put $s_n(x) = nx^2/(1 + nx^2)$. This sequence of continuous functions converges to 1 for all real $x \neq 0$ and to 0 for $x = 0$. What happens in the vicinity of $x = 0$? For $x = \alpha$ we consider different "variables with limit 0" and test $s_n(\alpha)$ for convergence. For this we must use Cauchy's concept of convergence.

Convergence of the s_n to a fixed limit s translates into "it is necessary and sufficient that, for infinitely large values of n, the sums s_n, s_{n+1}, s_{n+2}, *etc* ... should differ from the limit s, and thus from one another, by infinitely small magnitudes." Hence if n and n' are infinitely large numbers, then $s_n - s_{n'}$ must be an infinitely small magnitude. As Cauchy explained early on, an infinitely large number is a variable magnitude with limit $+\infty$.

Let's choose for our example a fixed, infinitely small magnitude $\alpha = m^{-1}$, m an infinitely large number. If we put $n = m^2$ and $n' = m$, then

$$s_n(\alpha) - s_{n'}(\alpha) = \frac{1}{2} - \frac{1}{m + 1},$$

and this is not infinitely small. Since the assumption of convergence in the vicinity of $x = 0$ is not fulfilled, we have no counterexample.

It is not surprising that Cauchy's theorems were thought to be false. First Abel, who greatly admired Cauchy's rigor in analysis, expressed scepticism. He did this in a footnote to his paper on the binomial series published in 1826, in which he also hinted at Fourier series. After the appearance of Robinson's book (1966), mathematicians have again tried to take seriously Cauchy's pronouncements about infinitely small magnitudes and about infinitely large values of the number n. By now there are scores of discussion papers. A comprehensive account was published by Bottazzini in 1992.

If one wants to understand these objectively existing obscurities, then it is not enough to look at just these two theorems. Rather, one must look at the overall development of Cauchy's mathematical views. It is hardly plausible that this great and versatile mathematician should have failed in the very first two theorems of his 1821 textbook. This becomes even more unlikely if we bear in mind that, as we shall see below, the second theorem, the one dealing with the continuity of the sum of a series, played a fundamental role in the further development.

In his many papers on mathematical physics and on Fourier analysis that date back to the years 1815–1820, Cauchy refrained from references to infinitely small

magnitudes and spoke only of the "very small." In the 1820s, frequently in the same context as before, he spoke about infinitely small numbers, and treated them as independent mathematical magnitudes, without recourse to the concept of a variable. In the first few years after 1816, while teaching at the Ecole Polytechnique, Cauchy avoided the use of the infinitely small. This provoked growing criticism on the part of his colleagues, including the physicist Petit, who emphasized the didactical and practical advantages of the use of infinitely small magnitudes. In 1819 and in 1820, the Conseil d'Instruction at the Ecole exerted strong pressure on Cauchy, but this alone would not have made this rather stubborn man change his mind. Around 1820, he must have realized that infinitesimal considerations were a powerful research method at a time when he was in a state of constant rivalry, especially with Poisson. Nevertheless, he insisted on using the derivative function instead of the quotient of infinitely small differentials, and assigned to infinitely small magnitudes a role linked to the concept of continuity.

In this connection, Cauchy said in the introductory remarks to his *Cours*: "En parlant de la continuité des fonctions, je n'ai pu me dispenser de faire connaître les propriétés principales des quantités infiniment petites, propriétés qui servent de base au calcul infinitésimal." Beginning in 1823, in introductions to books, he claimed consistently to have harmonized the intuitive appeal of infinitely small magnitudes with the rigor of analysis.

In the 1820s Cauchy was a very busy man. He taught not only at the Ecole Polytechnique but also at other institutions, and he published thousands of pages each year. The care with which these works were composed is admirable. Having recognized the usefulness of infinitely small magnitudes in research, he must have been tempted not only to use them in concurrently written textbooks but also to justify them rigorously. This was a stepwise process. At the beginning we still encounter the traditional locutions about variables with limit zero or infinity. In time, the infinitely small magnitudes, soon referred to as numbers, acquire independence and are handled like 'genuine' numbers. The year 1823 witnessed a breakthrough in the form of an abstract theory, to be discussed in Section 7.

We can look at the temporal sequence of Cauchy's books and concurrent research papers from the viewpoint of the genesis of a new theory. In this context, from which one must not dislodge individual theorems and subject them to critical inspection in isolation, a substantial portion of the historical evolution of mathematics becomes understandable.

6. CAUCHY AND THE BINOMIAL SERIES. In Chapter VI of his *Cours d'analyse* Cauchy treated the problem of the expansion, "to every possible extent," of the function $(1+x)^{\mu}$. After Newton, the binomial series had found many uses, and Euler and others had tried in vain to find a proof of its validity for arbitrary real μ. Bolzano's correct proof of 1816 was overlooked. Nowadays, we make use of Taylor's formula with remainder.

For x with $|x| < 1$ Cauchy defines

$$\Phi(\mu) = 1 + \frac{\mu}{1}x + \frac{\mu(\mu-1)}{2!}x^2 + \cdots$$

as a function of the real variable μ. He had proved earlier that this series converges and that the function satisfies the functional equation $\Phi(\mu)\cdot\Phi(\mu') = \Phi(\mu+\mu')$. The partial sums of the series are polynomials in μ, and thus

continuous functions. By Theorem I on series, $\Phi(\mu)$ is also a continuous function of μ.

As proved earlier, $(1+x)^{\mu}$ is continuous because it is an exponential function. Thus we have two continuous (nonzero) solutions of the functional equation. But Cauchy proved that there is only one continuous solution of this functional equation with $\Phi(0) = 1$. Hence $(1+x)^{\mu}$ is equal to the series.

As was pointed out by Bottazzini (1992), this proof can be viewed as a key that enables us to understand how Cauchy constructed his *Cours*. In his proof Cauchy may be said to be using everything treated earlier: continuity, convergence, Theorem I that links the two, and the treatment of functional equations. Cauchy's handling of the proof illustrates the technical superiority of his *conceptual* approach to the *algorithmic* methods of his predecessors. The same is true of the theorem on the intermediate value property. Cauchy proves it, in an appendix to his *Cours* (Note III), for continuous functions, whereas his predecessors, including Euler, failed to prove it even for the special case of real polynomials. The inference is drawn not from the algebraic property of being a polynomial but from the conceptual property of continuity. Bolzano noticed this already in 1817. Unlike Bolzano, Cauchy does not mention that he has managed to solve two 'much-courted' problems.

As a corollary, putting $\mu = 1/\alpha$ and replacing x by αx, with $|x| < 1/\alpha$, Cauchy obtains the series

$$(1+\alpha x)^{1/\alpha} = 1 + \frac{x}{1!} + \frac{x^2}{2!}(1-\alpha) + \frac{x^3}{3!}(1-\alpha)(1-2\alpha) + \cdots,$$

which yields the exponential series for α infinitely small. We recognize Euler's setup, but this time the proof is correct. The same holds for the second corollary, which deals with the series for the logarithm:

$$(1+x)^{\mu} = e^{\mu \log(1+x)} = 1 + \frac{\mu \log(1+x)}{1!} + \mu^2(\cdots)$$

yields

$$\lim_{\mu \to 0} \frac{(1+x)^{\mu} - 1}{\mu} = \log(1+x).$$

On the other hand, by the binomial theorem,

$$\frac{(1+x)^{\mu} - 1}{\mu} = x - \frac{x^2}{2}(1-\mu) + \frac{x^3}{3}(1-\mu)\left(1 - \frac{\mu}{2}\right) - + \cdots$$

For $\mu \to 0$ this yields the series for the logarithm. Those acquainted with Euler's *Introductio* will notice to what extent Cauchy reminds them of his predecessor. After all, the *Cours* is subtitled *Analyse algébrique*, and treats the circle of topics of the *Introductio* 'algebraically,' without applying the 'transcendental' methods of the differential and integral calculus.

7. CAUCHY'S CONCEPT OF NUMBER. Euler's operating with infinitely small and infinitely large numbers was not very problematic when it came to expressions and their algorithmic transformation in accordance with the ordinary (rational) rules of computation.

This is different in Cauchy's case, because Cauchy thought in *concepts*. If a function $f(x)$ was determined by its values $f(x_0)$ for all real values of the argument x_0, how was it to be determined for values $x_0 + \alpha$ with α infinitely small? The vague talk about variables with limit 0 had to be made precise if, as Cauchy did ever more frequently, one wanted to talk about infinitely small *numbers*. One gets the impression that Cauchy felt this need when he wrote his second textbook, the *Résumé* of 1823. He introduced his *system* of infinitely small numbers. We recall the formation of the real numbers and then talk about Cauchy's analogous, but more general, number concept.

Given the rational numbers, one of the possible ways of creating the reals is even connected with Cauchy's name. We form sequences of rational numbers and single out the Cauchy sequences (a_n): (a_n) is a Cauchy sequence if for every $\epsilon > 0$ there is an $N(\epsilon)$ such that $|a_n - a_m| < \epsilon$ for all $n, m \geq N(\epsilon)$. Every Cauchy sequence represents a real number, and two such sequences represent the same real number if and only if their differences form a null sequence. (One began to think in terms of equivalence relations around 1900; I am using here locutions found in the works of Heine and Cantor (1872).)

Cauchy can assume the real functions as given. A real function $f(u)$, defined for $u > 0$, represents an infinitely small number if $\lim f(u) = 0$ for $\lim u = 0$. Cauchy says that the function $g(u) = u$ represents the basis ω of his system and writes $f(\omega)$ for the number represented by $f(u)$. It is obvious that one computes with these numbers the way one computes with the functions that represent them. Cauchy mentioned his conception many times but did not advance it much. In particular, he provided no corresponding explanation for infinitely large numbers which, as we saw in Section 5, he used very often.

A consistent development of Cauchy's setup brings us to a general definition of Cauchy numbers (Laugwitz 1991):

Every real function $f(u)$ defined on an interval $0 < u < p$ represents a Cauchy number. Two such functions $f(u)$ and $g(u)$ represent the same Cauchy number if and only if there is an interval $0 < u < q$ in which $f(u) = g(u)$. We write ω for the Cauchy number represented by the identity function.

Operations and relations are defined for Cauchy numbers in a fairly obvious way by means of operations and relations for their representatives.

$f(\omega) = g(\omega)$ if and only if $f(u) = g(u)$ in an interval $0 < u < p$. Similarly, $f(\omega) > g(\omega)$ if and only if $f(u) > g(u)$ in an interval $0 < u < p$.

$\omega^{-1} > r$ for every real r; ω^{-1} is an infinitely large number. The real numbers are the Cauchy numbers represented by the constant functions $h(u) = r$.

Note that this introduction of the Cauchy numbers is analogous to the generation of the real numbers in terms of rational sequences, which are rational functions over the integers. When we work with real numbers we don't think of their genesis in terms of rational sequences. Similarly, once we get used to working with Cauchy numbers we don't have to think of their genesis in terms of real functions.

We can now tell how to obtain a real function $F(x)$ defined on the interval $x_0 \leq x \leq X$ for a Cauchy number $\beta = b(\omega)$ in this interval. $F(\beta) = F(b(\omega))$ is represented by $F(b(u))$. This makes sense, because for small $u > 0$ we have $x_0 \leq b(u) \leq X$. Without having to mention it every time, $F(\beta)$ is a well-defined Cauchy number.

The continuity property $F(x + \alpha) - F(x) \approx 0$ for $\alpha \approx 0$ is also set up in this way, with $\alpha \approx 0$ denoting infinitely small α.

Cauchy speaks of systems in the context of group theory as well. A group—this is a word he does not use—is a system whose elements are represented by, say, substitutions (i.e., permutations of a finite set). In modern terms, his systems of numbers are vector spaces, and even rings. This doesn't do much for us, but it explains why his formulation was too abstract for his contemporaries. This was also true of his and Galois' group theory.

Today group theory is described in the language of modern algebra. We can try to do this for the Cauchy numbers as well.

Consider the ring R of real functions defined in a right neighborhood of 0, and in this ring, the ideal I of functions that vanish in a right neighborhood of 0. Then R/I is the ring of Cauchy numbers. If we use the axiom of choice and extend I to a maximal ideal J, then R/J is a nonarchimedean ordered field that corresponds to the $^*\mathbb{R}$ of Robinson's nonstandard analysis. Obviously, there are important properties of R/I and R/J other than their algebraic and order properties. One such is the possibility of continuing real functions from $\mathbb{R}$ to these extension rings.

My paper of 1991 contains a number of applications of these concepts. Here I'll limit myself to showing the logical equivalence of the two definitions of continuity.

Let $F(x)$ be a real function defined in a neighborhood of x_0. The two definitions are:

(I) F is $\epsilon - \delta$-continuous at x_0 if for every $\epsilon > 0$ there is a $\delta > 0$ such that $|F(x) - F(x_0)| < \epsilon$ for all real x with $|x - x_0| < \delta$.

(II) F is C-continuous at x_0 if $F(x_0 + \alpha) - F(x_0) \approx 0$ for all $\alpha \approx 0$.

(I) implies (II): Let α be represented by $a(u)$. Since $\alpha \approx 0$, it follows that $|\alpha| < \delta$, and thus also $|a(u)| < \delta$ for small $u > 0$. But then $|F(x_0 + a(u)) - F(x_0)| < \epsilon$ for small $u > 0$, and so $|F(x_0 + \alpha) - F(x_0)| < \epsilon$ for every positive real ϵ. (II) follows.

(II) implies (I): Proof by contradiction. Suppose that (I) is false. Then there is an exceptional ϵ, say $p > 0$, for which there is no corresponding δ. In particular, for every $u > 0$ there must be an $a(u)$, $0 < a(u) < u$, with $|F(x_0 + a(u)) - F(x_0)| \geq p$.

For $\alpha = a(\omega)$ we have a contradiction to the assumption (II).

The Euler-Cauchy convergence condition $s_n - s \approx 0$ for all infinitely large n is now a provable theorem. If it holds, then the ϵ-condition follows. Otherwise there is an exceptional ϵ, say $p > 0$, such that for every $u > 0$ there is an $N(u) > 1/u$ with $|s_{N(\omega)} - s| \geq p$, and thus not infinitely small. The converse is obvious.

8. DIFFERENTIAL AND INTEGRAL. The account of the foundations of the differential and integral calculus, as we know them, goes back to Cauchy who, of course, presented them in terms of infinitesimal mathematics and not in terms of epsilontics. Given its historical importance, we reproduce it in brief form.

If a function $f'(x)$ satisfies the condition

$$\frac{f(x+\alpha) - f(x)}{\alpha} \approx f'(x)$$

for all infinitely small α, then it is called the derivative of $f(x)$. All rules follow easily, especially if we write

$$f(x+\alpha) = f(x) + f'(x)\alpha + o(x,\alpha)\alpha \text{ with } o(x,\alpha) \approx 0.$$

Cauchy views the differential dy of $y = f(x)$ as a function of two variables, $dy = f'(x)\,dx$, which is linear with respect to the real variable dx.[3]

In his *Résumé* of 1823 Cauchy treated the differential and integral calculus. It is noteworthy that he proved the integrability of continuous functions. It is possible that he was motivated by the desire to bring continuity into prominence. In the 18th century, integrals were viewed as antiderivatives, and this no longer satisfied the new needs of mathematical physics that required the integration of functions for which no antiderivative was known.

Let $f(x)$ be continuous on the interval $a \leq x \leq b$, and let $a = x_0 < x_1 < \cdots < x_n < x_{n+1} < \cdots < x_N = b$ be a subdivision with $\alpha_n = x_n - x_{n-1} \approx 0$. All the numbers in the n-th subinterval are infinitely close to the same real number and, because of the continuity of $f(x)$, the corresponding functional values differ by infinitely little from the functional value at that point, so that there are numbers M_n and m_n with $M_n - m_n \approx 0$ and $M_n \geq f(x) \geq m_n$ in the n-th subinterval. Using the concepts introduced by Darboux only around 1870, we can define the upper integral of f as the infimum of all upper sums and the lower integral as the supremum of all lower sums. (Cauchy used intermediate sums.) It is clear that no lower sum is greater than an upper sum. If we can show that the upper sum $\Sigma M_n \alpha_n$ and the lower sum $\Sigma m_n \alpha_n$ associated with our infinitesimal subdivision differ by infinitely little, then both are infinitely close to the same real number that is the definite integral $\int_a^b f(x)\,dx$.

For every real $p > 0$ we have $0 \leq M_n - m_n < p$. Hence

$$0 \leq \sum M_n \alpha_n - \sum m_n \alpha_n = \sum (M_n - m_n)\alpha_n < p \sum \alpha_n = p(b - a).$$

Thus the difference between the upper sum and lower sum is indeed infinitely small.

This very intuitive proof, like Cauchy's own, avoids the explicit use of the uniformity of continuity.

9. FOURIER SERIES OF CONTINUOUS FUNCTIONS. Having obtained the fundamental theorems on differentiation and integration one can forget the way they were derived, because all that counts now is these theorems. That is why the infinitely small numbers retained the right to exist only in physics and in differential geometry, e.g., for the derivation of differential equations from the consideration of infinitely small elements of space and time.

But if one had a tidy theory of infinitely small numbers, as Cauchy did, one could also use it in other ways, and this was done in Paris around 1820. As a relevant example we consider Fourier series. The argument that follows is found not only in the works of Cauchy and Poisson but also in a surviving manuscript of

[3]Besides, the derivative function, as defined by Cauchy, is continuous in its own right. Indeed, if we make the variable substitution $z = x + \alpha$ and $\beta = -\alpha$, then we have:

$$f'(z) \approx \frac{f(z + \beta)}{\beta} = \frac{f(x + \alpha)}{\alpha} \approx f'(x),$$

so that $f'(x + \alpha) \approx f'(x)$ for every infinitely small α.

A school mathematician may be shocked that all derivative functions are continuous, but in research in the area of modern analysis one almost always assumes continuous differentiability. This property is logically equivalent to uniform differentiability: $f'(z) = \lim(f(x + h) - f(x + k)/(h - k)$ for $h, k \to 0$.

Gauss that probably dates back to 1815. After a simple transformation, the Euler series in para. 4 yields

$$\text{(A)} \qquad 1/2 + q\cos x + q^2\cos 2x + q^3\cos 3x + \cdots + q^n\cos nx + \cdots$$
$$= \frac{1-q^2}{1-2q\cos x + q^2} =: \pi\delta(x).$$

For $\epsilon = 1 - q \approx 0$, $0 < \omega \le x \le 2\pi - \omega$, $\omega = \sqrt[3]{\epsilon} \approx 0$, we get, with Cauchy,

$$0 < \frac{1-q^2}{1-2q\cos x+q^2} = \frac{\epsilon(1+q)}{1-2q(1-2\sin^2 x/2)+q^2} = \frac{\epsilon(1+q)}{\epsilon^2+4q\sin^2 x/2}$$
$$< \frac{\epsilon}{\sin^2 x/2} \le \frac{\epsilon}{\sin^2 \omega/2} \approx 0.$$

The function $\delta(x)$ defined in (A) is 2π-periodic, positive everywhere, and infinitely small outside of intervals of infinitely small length around the values $x = 2\pi g$ with integral g, and its integral over an interval of length 2π is 1. (The latter is obtained by termwise integration of the series on the left. That this is permissible can be seen by considering the majorizing geometric progression.)

Today we call such a function a Dirac delta function. The delta functions, outlawed in standard real analysis, turn out to be magic wands in analysis and in physics. One knew this around 1820, but then the whole thing was forgotten.

If $f(t)$ is continuous and periodic with period 2π, then

$$\text{(B)} \qquad f(t) \approx \int \delta(x) f(x+t)\,dx = \int \delta(x-t) f(x)\,dx,$$

where the integrations extend over any interval of length 2π. Indeed, if $g(x)$ is continuous and periodic with period 2π, then

$$\int_{-\pi}^{+\pi} \delta(x) g(x)\,dx = \int_{-\pi}^{-\omega} + \int_{-\omega}^{+\omega} + \int_{+\omega}^{+\pi}.$$

Here the first and last integrals on the right are infinitely small because δ is. Since $\int_{-\omega}^{+\omega} \delta(x)\,dx \approx 1$ and g is continuous, the mean value theorem assigns to the middle integral a value that is infinitely close to $g(0)$.

Now we apply this to

$$\delta(x-t) = \frac{1}{2\pi} + \frac{1}{\pi}\sum_{n\ge 1} q^n \cos n(x-t)$$
$$= \frac{1}{2\pi} + \frac{1}{\pi}\sum_{n\ge 1} q^n[\cos nx\cos nt + \sin nx \sin nt].$$

For the same reason as before, we can integrate termwise. Using the Fourier coefficients

$$a_n = \frac{1}{\pi}\int_{-\pi}^{+\pi} f(x)\cos nx\,dx, \qquad b_n = \frac{1}{\pi}\int_{-\pi}^{+\pi} f(x)\sin nx\,dx,$$

we get from (B)

$$\text{(C)} \qquad f(t) \approx \frac{a_0}{2} + \sum_{n\ge 1} q^n(a_n\cos nt + b_n \sin nt)$$

for continuous 2π-periodic $f(t)$ and $1 > q \approx 1$.

Unfortunately, we cannot subsequently put $q = 1$: this would yield a Fourier series. Since 1873 we have known that continuity of a function does not suffice for it to be equal to its Fourier series. This result of du Bois-Reymond came as a surprise. In the unpublished manuscript mentioned earlier, Gauss drew the false conclusion that it does. Cauchy was more careful. He said that the Fourier series

$$\text{(D)} \qquad \frac{a_0}{2} + \sum_{n \geq 1} a_n \cos nt + b_n \sin nt$$

is equal to the function if both series (C) and (D) converge. The reader can verify this.

BIBLIOGRAPHY

Becker, O. 1954 *Grundlagen der Mathematik in geschichtlicher Entwicklung*. Verlag Karl Alber, Freiburg und München.

Bottazzini, U. 1992 See Cauchy 1821/1992.

Cauchy, A. L. 1821/1992 *Cours d'analyse de l'Ecole polytechnique*, Paris. œuvres ser. 2, v. 3. Reprint, with an introduction by U. Bottazzini, CLUEB, Bologna.

Cauchy, A. L. 1823 *Résumé des leçons sur le calcul infinitesimal*, Paris. œuvres ser. 2, v. 4.

Euler, L. 1748/1983 *Introductio in analysin infinitorum*. Lausanne. Opera omnia ser. 1, v. 8; German translation by H. Maser with an introduction by W. Walter: *Einführung in die Analysis des Unendlichen*. Springer, Berlin.

Euler, L. 1983 *Leonhard Euler 1707–1783*. Beiträge zu Leben und Werk. *Gedenkband des Kantons Basel-Stadt. Birkhäuser, Basel.*

Fellmann, E. 1995 *Leonhard Euler*. Rowohlts Monographien, Basel.

Laugwitz, D. 1986/1994 *Zahlen und Kontinuum. Eine Einführung in die Infinitesimalmathematik* B. I. Wissenschaftsverlag, Mannheim.

Laugwitz, D. 1991 Cauchy-Zahlen als Grundlage der Infinitesimalmathematik. *Mathematische Semcster-berichte* **38**, 175–213.

L'Hospital, G. F. A. 1696 *Analyse des infiniment petits*. Paris.

Robinson, A. 1966 *Nonstandard Analysis*. North Holland, Amsterdam. (Paperback, with an introduction by W. A. J. Luxemburg: Princeton University Press, 1995.)

Ahornweg 23
D-64367 Mühltal, Germany

More Mathematical Double Entendres:

Open Set: Something that Pete Sampras usually wins at Forest Hills
Arcsine: What Noah used to direct the animals to his boat
Inaccessible Cardinal: A VIP at the Vatican
Cyclic Group: A bunch of people in leather jackets riding Harleys
Compact Manifold: A part of the engine of a Honda Civic
Splitting Field: Something to look out for near the San Andreas fault
Mean Deviation: An activity involving especially nasty behavior

Contributed by David Sprows, Villanova University

Q: What is $|\text{Rogers} - x| < \varepsilon$?

A: Mr. Rogers' neighborhood.

Contributed by Eric Key, University of Wisconsin–Milwaukee

PROBLEMS AND SOLUTIONS

Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**

with the collaboration of Paul T. Bateman, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Underwood Dudley, Michael A. Filaseta, Ira M. Gessel, Bart Goddard, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Robert Israel, Murray S. Klamkin, Daniel J. Kleitman, Fred Kochman, Frederick W. Luttmann, Frank B. Miles, Richard Pfiefer, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before February 28, 1998; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.*

PROBLEMS

10606. *Proposed by Thomas Zaslavsky, Binghamton University, Binghamton, NY.* Given a positive integer m, show that there is a positive integer n such that, for any group G of order at least n, it is possible to choose m elements $g_1, g_2, \ldots, g_m$ of G so that no product of the form $g_{i_1}^{\pm 1} g_{i_2}^{\pm 1} \cdots g_{i_k}^{\pm 1}$ with $1 \le k \le m$ and distinct subscripts $i_1, i_2, \ldots, i_k$ in $\{1, 2, \ldots, m\}$ equals the identity.

10607. *Proposed by Juan-Bosco Romero Márquez, Universidad de Valladolid, Valladolid, Spain.* Evaluate

$$\lim_{n\to\infty} \left(\frac{2^x + 4^x + \cdots + (2n)^x}{1^x + 3^x + \cdots + (2n-1)^x} \right)^n$$

for $x > 0$.

10608. *Proposed by Victor Zalgaller, Steklov Mathematical Institute, St. Petersburg, Russia.* Let S be a compact convex set in the plane. If l is any line of support for S, let $f(l)$ be the length of the shortest curve that begins and ends on l and that together with l surrounds S. Prove that if $f(l)$ is independent of l, then S is a circle.

10609. *Proposed by Donald E. Knuth, Stanford University, Stanford, CA.* Let

$$a(l, m, n) = \sum\nolimits_{k=0}^{l} \binom{n}{k} (l + m - k)^{n-k} (k - l)^k.$$

Prove that

$$\sum\nolimits_{l=1}^{n} a(l, m, n) = \frac{m+n+1}{2} a(n, m, n) - \frac{m+1}{2} m^n.$$

10610. *Proposed by Richard Hall, University of Portsmouth, Portsmouth, England.* Given a positive integer m, let $C(m)$ be the greatest positive integer k such that, for some set S of m integers, every integer from 1 to k belongs to S or is a sum of two not necessarily distinct elements of S. For example, $C(3) = 8$ with $S = \{1, 3, 4\}$.

(a) Show that, for all $\epsilon > 0$, $1/4 < C(m)/m^2 < 1/2 + \epsilon$ for all sufficiently large m.

(b)* Improve the asymptotic bounds in part **(a)**.

10611. *Proposed by Zoltán Sasvári, Technical University of Dresden, Dresden, Germany.* Find the largest value of a and the smallest value of b for which the inequalities

$$\frac{1+\sqrt{1-e^{-ax^2}}}{2} < \Phi(x) < \frac{1+\sqrt{1-e^{-bx^2}}}{2}$$

hold for all $x > 0$, where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy.$$

10612. *Proposed by John P. Robertson, Anistics/Aon, New York, NY.* Fermat proved that there are no nontrivial 4-term arithmetic progressions all of whose terms are integer squares.
(a) Find all 5-term arithmetic progressions such that all terms but the fourth are squares.
(b) Call two arithmetic progressions *essentially different* if the ratios of corresponding terms differ. For each integer $m \geq 6$ show that there are infinitely many essentially different m-term arithmetic progressions such that the first 3 terms and the mth term are squares.

SOLUTIONS

A Fairly General Family of Integrals

10393 [1994, 573]. *Proposed by Jean Anglesio, Garches, France.* Show that

$$\int_0^\infty \frac{e^{-ax}\left(1-e^{-x}\right)^n}{x^r}\, dx = \frac{(-1)^r}{(r-1)!} \sum_{k=0}^{n} \binom{n}{k} (-1)^k (a+k)^{r-1} \log(a+k)$$

where $a \geq 0$ and $1 \leq r \leq n$ (except for $a = 0, r = 1$).

Solution by Jet Wimp, Drexel University, Philadelphia, PA. The Gamma function $\Gamma(\sigma)$ is analytic for $\sigma \neq 0, -1, -2, \ldots$; it satisfies $\Gamma(\sigma+1) = \sigma\Gamma(\sigma)$, so $\Gamma(r) = (r-1)!$ when r is a positive integer. Also, we have

$$\begin{aligned}
\lim_{\sigma\to 1-r} \frac{d}{d\sigma}\frac{1}{\Gamma(\sigma)} &= \lim_{\sigma\to 1-r} \frac{d}{d\sigma}\left((\sigma+r-1)\frac{\sigma(\sigma+1)\ldots(\sigma+r-2)}{\Gamma(\sigma+r)}\right) \\
&= \lim_{\sigma\to 1-r}\left((\sigma+r-1)\frac{d}{d\sigma}\left(\frac{\sigma(\sigma+1)\ldots(\sigma+r-2)}{\Gamma(\sigma+r)}\right) + \frac{\sigma(\sigma+1)\ldots(\sigma+r-2)}{\Gamma(\sigma+r)}\right) \\
&= (-1)^{r-1}(r-1)! \qquad (A)
\end{aligned}$$

for $r = 1, 2, \ldots$.

Expanding $\left(1-e^{-x}\right)^n$ by the binomial theorem, integrating term by term, and using $\int_0^\infty e^{-px}x^{\sigma-1}\,dx = \Gamma(\sigma)p^{-\sigma}$ for $p, \sigma > 0$ yields

$$\int_0^\infty e^{-ax}\left(1-e^{-x}\right)^n x^{\sigma-1}\, dx = \frac{\sum_{k=0}^{n}\binom{n}{k}(-1)^k(a+k)^{-\sigma}}{1/\Gamma(\sigma)} \qquad (B)$$

for $a > 0$. The range of validity of the integral in (B) may be extended from $\sigma > 0$ to $\sigma > -n$. In fact, this integral is analytic for complex σ with real part greater than $-n$.

For $\sigma = 1-r, r = 1, 2, \ldots, n$, the right side of (B) is indeterminate: the sum in the numerator is zero (it is an nth difference of $a+k$ to a nonnegative integral power less than n) and the denominator must be zero in the limit since the integral is clearly not. We now take the limit of (B) as $\sigma \to 1-r, r = 1, 2, \ldots, n$ using L'Hospital's rule. Differentiating the numerator using $\frac{d}{d\sigma}(a+k)^{-\sigma} = -(a+k)^{-\sigma}\ln(a+k)$ and the denominator using (A) gives the result.

For other values of σ for which the integral exists, its value is given by the formula (B). Extension of the result to $a = 0$ for values of r for which the integral converges, i.e., $r = 2, \dots, n$, is trivial.

Editorial comment. Michael Vowe noted that the cases $r = 1, 2, 3$ can be found in I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, prepared by A. Jeffrey, Academic Press, 1980 (formulas 3.411/19 p. 377, 3.411/20 p. 377, and 3.443/2 p. 386). He also noted that the case of all $r < n$ is contained in a general formula in section 167 of T. J. Bromwich, *An Introduction to the Theory of Infinite Series*, Second Edition, Macmillan, 1926, p. 473. Murray S. Klamkin reported that the case $r = n$ appears in *Pi Mu Epsilon Journal*, Problem 805 [1993, 545; 1994, 703] and extended the method of J. S. Frame's solution of that problem (integration by parts, with a careful consideration of the behavior at the endpoints) to cover all cases of this problem.

Solved also by D. Borwein (Canada), P. Bracken (Canada), D. Bradley, E. Braune (Austria), D. Callan, R. J. Chapman (U. K.), H. Chen, J. L. Collett (U. K.), L. Euler (transmitted by D. Zeilberger), P. Flajolet (France), M. Golomb, S. A. Greenspan, J.-P. Grivaux (France), H. van Haeringen (The Netherlands), R. Holzsager, G. L. Isaacs, F.-A. Izadi (Iran), G. Keselman, P. Khajeh-Khalili, M. S. Klamkin (Canada), M. J. Knight, O. P. Lossers (The Netherlands), K. McInturff, K. D. McLenithan, A. Pechtl (Germany), A. Pedersen (Denmark), H. Prodinger (Austria), M. Vowe (Switzerland), T. White, A. N. 't Woord (The Netherlands), K. Zacharias (Germany), P. J. Zwier, NSA Problems Group, Prague Problem Solution Group (Czech Republic), WMC Problems Group, and the proposer.

A Persistent Distribution

10394 [1994, 575]. *Proposed by Ignacy I. Kotlarski, Oklahoma State University, Stillwater, OK.* Let $N, Z_1, Z_2, \dots$ be a sequence of independent random variables, where N follows the geometric distribution with $\mathrm{Prob}(N = n) = p(1-p)^{n-1}$ for $n = 1, 2, \dots$ with $0 < p < 1$, while the Z_j are identically distributed complex random variables $Z_j = X_j + iY_j$ where (X_j, Y_j) have density

$$f_{Z_j}(x, y) = \begin{cases} \dfrac{a}{2\pi}\,|z|^{a-2} & \text{for } |z| < 1 \\ 0 & \text{for } |z| \geq 1 \end{cases}$$

where $z = x + iy$ and $a > 0$. Find the distribution of $W = Z_1 \cdot Z_2 \cdots Z_N$.

Solution by David Callan, University of Wisconsin, Madison, WI. W has the same distribution as the Z_j with a replaced by pa.

Switching to polar coordinates, $Z_j = X_j + iY_j = R_j e^{i\Theta_j}$ where R_j and Θ_j are independent, Θ_j is uniform on $(0, 2\pi)$, and R_j has the power density with parameter a: $f(r) = ar^{a-1}$ for $0 < r < 1$. Thus $W = Re^{i\Theta}$ where $R = R_1 R_2 \cdots R_N$ and $\Theta = \Theta_1 + \Theta_2 + \cdots + \Theta_N$ reduced modulo 2π and R, Θ are independent.

It is clear that the distribution of Θ is uniform on $(0, 2\pi)$. Indeed, if U, V are independent random variables with V uniform on $(0, 2\pi)$ then $W = U + V$ reduced modulo 2π is also uniform on $(0, 2\pi)$ regardless of the distribution of U. Thus, for any fixed k, $\Theta_1 + \Theta_2 + \cdots + \Theta_k$ reduced modulo 2π is uniform on $(0, 2\pi)$, and so Θ is also.

As for the distribution of R, let $U_j = -\log R_j$. Then U_j has the exponential density with parameter a: $g(u) = ae^{-au}$ for $u > 0$, whose moment generating function is $\psi(t) = \sum_{n \geq 0} \mathrm{E}(U_j^n)t^n/n! = a/(a-t)$. Hence the moment generating function of $U = U_1 + U_2 + \ldots + U_N$ is

$$\sum_{n \geq 1} \mathrm{Prob}(N = n)\psi(t)^n = p\psi(t) \sum_{n \geq 1} (1-p)^{n-1}\psi(t)^{n-1} = \frac{pa}{pa - t}.$$

Thus U has exponential density with parameter pa and so $R = e^{-U}$ has the power density with this parameter.

Solved also by R. A. Agnew, N. Bouzar, W. J. Buhler (Germany), R. Ehrenborg (Canada), J. A. Grzesik, V. Hernández (Spain), S. J. Herschkorn, R. Holzsager, G. Keselman, J. H. Lindsey II, O. P. Lossers (The Netherlands), D. K. Nester, T. Shore & D. B. Tyler, T. White, A. N. 't Woord (The Netherlands), and the proposer.

Winding Around Points in a Chessboard

10401 [1994, 682]. *Proposed by Donald E. Knuth, Stanford University, Stanford, CA.* A closed *knight's tour* of an m by n chessboard is a sequence $\langle (x_k, y_k) \rangle$ for $0 \le k < mn$ such that each pair of integers (x, y) with $0 \le x < m$ and $0 \le y < n$ occurs exactly once in the sequence, and $(x_{k+1} - x_k)^2 + (y_{k+1} - y_k)^2 = 5$ for all k (including $k = mn - 1$ with $(x_{mn}, y_{mn}) = (x_0, y_0)$). Such a tour defines a closed *contour* C if we connect adjacent points (x_k, y_k) and (x_{k+1}, y_{k+1}) with straight line segments.

Let w_{ij} be the winding number of C about the point $(i - \frac{1}{2}, j - \frac{1}{2})$. Prove that

$$\sum_{k=0}^{mn-1} (x_k y_{k+1} - x_{k+1} y_k) = 2 \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} w_{ij}.$$

Solution by Richard Stong, Rice University, Houston, TX. Let $A_{\pm}(C) = \sum_{k=0}^{mn-1} (x_k y_{k+1} - x_{k+1} y_k)$. Note that $A_{\pm}(C)$ can be interpreted as the signed area inside the contour C (a point P being considered as inside C with multiplicity the winding number of C around P). We show the following more general fact.

Proposition. *Let C be a polygonal closed contour with vertices $(x_0, y_0), (x_1, y_1), \ldots, (x_{N-1}, y_{N-1}), (x_N, y_N) = (x_0, y_0)$ such that all the vertices have integer entries and no edge of C goes through any point of the form $(i - 1/2, j - 1/2)$ for integer i and j. Let $w_{ij}(C)$ denote the winding number of C about the point $(i - 1/2, j - 1/2)$. Then*

$$A_{\pm}(C) = \sum_i \sum_j w_{ij}(C). \qquad (*)$$

Proof. First note that both sides of the equation $(*)$ are additive under the following basic operation: take the disjoint union of two closed contours C_1 and C_2 and remove any edges that cancel. Next note that $(*)$ holds if C is the standard contour around a unit square (with vertices $(a, b), (a + 1, b), (a + 1, b + 1), (a, b + 1), (a, b)$), since the area is 1 and the winding number $w_{a+1,b+1}(C) = 1$ and all other $w_{ij}(C) = 0$. By symmetry, $(*)$ also holds for the reverse of this contour. Thus, by repeated use of the basic operation, we see that $(*)$ holds if all edges of C are parallel to the coordinate axes. Also note that $(*)$ holds for the triangular contour C with vertices $(0, 0), (a, 0), (0, b), (0, 0)$ with $a > 0$, $b > 0$ and exactly one of a and b even (so that no point of the form $(i - 1/2, j - 1/2)$ lies on the edge from $(a, 0)$ to $(0, b)$). For this contour, the area is $ab/2$, and it has winding number 1 about the $ab/2$ points of the form $(i - 1/2, j - 1/2)$ in its interior. By symmetry $(*)$ also holds for the contours obtained from C by the rigid motions of the plane that preserve the integer lattice. Repeated use of the basic operation shows that $(*)$ holds for all contours. □

The proposition holds for all polygonal closed contours with integral vertices if one interprets the winding number about points of the form $(i - 1/2, j - 1/2)$ on the boundary correctly. The method of proof suggests that this result is a form of Pick's theorem. Indeed, it can be obtained by applying Theorem 1 of B. Grünbaum and G. C. Shephard, Pick's Theorem, this MONTHLY 100 (1993) 150–161 to the integer lattice and the lattice consisting of integer points and points of the form $(i - 1/2, j - 1/2)$.

Solved also by E. Fernández Moral (Spain), M. Hoffman, R. Holzsager, J. H. Lindsey II, O. P. Lossers (The Netherlands), A. Nijenhuis, NSA Problems Group, and the proposer.

An Improper Double Integral

10411 [1994, 911]. *Proposed by Gord Sinnamon, University of Western Ontario, London, Ontario, Canada.* Let R be the region inside the unit circle and above the line $x + y = 1$. Calculate

$$\iint_R \frac{1}{(\log x)^2 + (\log y)^2} \frac{dx\,dy}{xy}. \qquad (*)$$

Solution I by Thomas M. McDonald, Gannon University, Erie, PA. The value of $(*)$ is $\pi \ln 2/2$. We show that if $0 < a < b$ and R is the first quadrant region bounded by the curves $x^a + y^a = 1$ and $x^b + y^b = 1$, then the value of $(*)$ is $(\pi/2)\ln(b/a)$.

The substitutions $u = \ln x$ and $v = \ln y$ transform $(*)$ to $\iint_S du\,dv/(u^2+v^2)$, where S is the third quadrant region between the curves $e^{bu} + e^{bv} = 1$ and $e^{au} + e^{av} = 1$. The substitutions $u = r\cos\theta$ and $v = r\sin\theta$ give $\iint_T (1/r^2)\, r\,dr\,d\theta$, where T is the region between the curves $e^{br\cos\theta} + e^{br\sin\theta} = 1$ and $e^{ar\cos\theta} + e^{ar\sin\theta} = 1$. If $f(\theta)$ is the function defined implicitly by the equation $e^{f(\theta)\cos\theta} + e^{f(\theta)\sin\theta} = 1$, then the required value is

$$\int_{\pi}^{3\pi/2}\int_{f(\theta)/b}^{f(\theta)/a} \frac{1}{r}\,dr\,d\theta = \int_{\pi}^{3\pi/2}\left(\ln\frac{f(\theta)}{a} - \ln\frac{f(\theta)}{b}\right)d\theta = \int_{\pi}^{3\pi/2} \ln\frac{b}{a}\,d\theta = \frac{\pi}{2}\ln\frac{b}{a}.$$

All of these manipulations are justified because each integrand is nonnegative and the final expression converges.

Solution II by O. P. Lossers, University of Technology, Eindhoven, The Netherlands. Because

$$\frac{\partial}{\partial y}\arctan\frac{\log x}{\log y} = -\frac{\log x}{y(\log^2 x + \log^2 y)},$$

$(*)$ equals

$$\lim_{\delta\downarrow 0}\int_{x=\delta}^{1}\frac{dx}{x\,\log x}\left[-\arctan\frac{\log x}{\log y}\right]_{y=1-x}^{\sqrt{1-x^2}} =$$

$$= \lim_{\delta\downarrow 0}\int_{\delta}^{1}\frac{dx}{x\,\log x}\left\{\arctan\frac{\log x}{\log(1-x)} - \arctan\frac{\log x^2}{\log(1-x^2)}\right\}.$$

The substitution $x = z^2$ reduces $\int_{x=\delta^2}^{1}\frac{dx}{x\,\log x}\arctan\frac{\log x}{\log(1-x)}$ to

$$\int_{z=\delta}^{1}\frac{2z\,dz}{z^2 2\log z}\arctan\frac{\log z^2}{\log(1-z^2)}\quad .$$

Hence, $(*)$ is found to be

$$\lim_{\delta\downarrow 0}\int_{x=\delta}^{\delta^2}\frac{dx}{x\,\log x}\arctan\frac{\log x}{\log(1-x)} = \tfrac{1}{2}\pi\log 2,$$

since, for some ε with $\delta^2 < \varepsilon < \delta$, we have

$$\int_{\delta}^{\delta^2}\frac{dx}{x\,\log x}\arctan\frac{\log x}{\log(1-x)} = \arctan\frac{\log\varepsilon}{\log(1-\varepsilon)}\int_{\delta}^{\delta^2}\frac{dx}{x\,\log x}.$$

The generalization in Solution I can be obtained by the same method.

Solution III by Jet Wimp, Drexel University, Philadelphia, PA. Define a continuous function on $[0,1]$ by $g(0) = 0$, $g(1) = \pi/2$, and $g(t) = \arctan(\ln(1-t)/\ln t)$ for $0 < t < 1$. We have

$$I = \iint_R \frac{dx\,dy}{xy\big((\log x)^2 + (\log y)^2\big)} = \int_0^1 \frac{dx}{x}\int_{1-x}^{\sqrt{1-x^2}}\frac{dy}{y\big((\log x)^2 + (\log y)^2\big)}$$

$$= \int_0^1 \frac{1}{x\ln x}\left[\arctan\left(\frac{\ln y}{\ln x}\right)\right]_{1-x}^{\sqrt{1-x^2}} dx = \int_0^1 \frac{g(x^2) - g(x)}{x\ln x}\,dx$$

$$= -\int_0^1 \frac{dx}{x\ln x}\int_{x^2}^{x} g'(t)\,dt.$$

Interchanging the order of integration gives

$$I = -\int_0^1 g'(t)\,dt \int_t^{\sqrt{t}} \frac{dx}{x\ln x} = -\int_0^1 g'(t)\Big[\ln|\ln x|\Big]_t^{\sqrt{t}}\,dt$$
$$= \ln 2\int_0^1 g'(t)\,dt = \ln 2\big(g(1)-g(0)\big) = \frac{\pi}{2}\ln 2.$$

Solved also by J. Anglesio (France), R. Bagby, R. J. Chapman (U. K.), D. A. Darling, J. S. Frame, M. Golomb, J. A. Grzesik, M. Hoffman, R. Holzsager, W. C. Lang, T. L. McCoy, D. K. Nester, A. Nijenhuis, T. Schonbek, R. Stong, D. Swearingen, A. A. Tarabay (Lebanon), M. Vowe (Switzerland), C. Y. Yildirim (Turkey), K. Zacharias (Germany), Anchorage Math Solutions Group, NSA Problems Group, and the proposer.

Balanced Sequences

10430 [1995, 71]. *Proposed by Fred Galvin, University of Kansas, Lawrence, KS, and John Isbell, SUNY, Buffalo, NY.* Let $D(a_1,\dots,a_k)$ denote the sum of the absolute deviations of the real numbers $a_1,\dots,a_k$ from their median. Call a sequence *balanced* if the $n-1$ quantities $D(a_1,\dots,a_k)+D(a_{k+1},\dots,a_n)$, $k=1,2,\dots,n-1$ are all equal.
(**a**) Show that, for each integer $n>1$, a nonconstant balanced sequence of n terms exists and is unique up to an affine transformation.
(**b**) Characterize the positive integers n for which there exists a *strictly increasing* balanced sequence of n terms.

Solution by Robin J. Chapman, University of Exeter, Exeter, UK. Statement (**a**) is false as given but becomes true when we consider only nondecreasing sequences. Let $b_j = a_{j+1}-a_j$ for $1\le j<n$. For part (**a**), it suffices to show that these differences are uniquely determined up to a constant factor and that there is a nontrivial nonnegative solution. For part (**b**), we show that all b_k's are nonzero if and only if $n\in\{1,2,4\}$ or n is a prime such that -1 and 2 generate the group of nonzero residues modulo n.

Let $d_k = D(a_1,\dots,a_k)+D(a_{k+1},\dots,a_n)$ for $1\le k<n$. The statement is trivial for $n\le 2$, so we assume that $n\ge 3$. We first consider odd n; let $r=(n-1)/2$ with $r\ge 1$. For $1\le j\le r$, we have

$$d_{2j-1} = -\sum_{i=1}^{j-1} a_i + \sum_{i=j+1}^{2j-1} a_i - \sum_{i=2j}^{j+r} a_i + \sum_{i=j+r+1}^{2r+1} a_i;$$
$$d_{2j} = -\sum_{i=1}^{j} a_i + \sum_{i=j+1}^{2j} a_i - \sum_{i=2j+1}^{j+r} a_i + \sum_{i=j+r+2}^{2r+1} a_i.$$

Equality between successive quantities in $d_1,\dots,d_{2r}$ yields

$$\begin{array}{lll} A_j: & a_j + a_{r+j+1} = 2a_{2j} & \text{for } 1\le j\le r; \\ B_j: & a_{j+1}+a_{r+j+1} = 2a_{2j+1} & \text{for } 1\le j\le r-1. \end{array}$$

Comparing neighboring equations in the sequence $A_1, B_1, A_2, \dots, A_{r-1}, B_{r-1}, A_r$ yields $b_j = 2b_{2j}$ and $b_{r+j+1} = 2b_{2j+1}$ for $1\le j\le r-1$. For $1\le k\le 2r$ and $k\notin\{r,r+1\}$, this yields $b_k = 2b_{(2k)}$, where (u) denotes the least positive residue of the integer u modulo n. Together with A_1, which is equivalent to $b_1=\sum_{i=2}^{r+1} b_i$, these relations form a complete set of equations for $\{b_j\}$. If $(2^s k)\notin\{r,r+1\}$ for all positive s, then $b_k = 2^S b_k$ when S is the order of 2 modulo n. It follows that $b_k=0$ unless k is congruent to $\pm 2^t$ modulo n for some t. When n is not prime, and also when n is prime but 2 and -1 do not generate the group of invertible residues modulo n, we obtain $b_k=0$ for some k, and the resulting sequence $\{a_k\}$ is not strictly increasing. If T is the least positive integer such that $2^T\equiv\pm1 \pmod n$, then $b_{(2^x)} = 2^{T-x-1}b$ and $b_{(-2^x)} = 2^{T-x-1}b'$, where $\{b,b'\}=\{b_r,b_{r+1}\}$. The relation

$b_1 = \sum_{i=2}^{r+1} b_i$ now becomes $2^{T-1}b = ub + vb'$, where u and v are positive integers with $u + v = 2^{T-1}$. Hence $b = b'$, and so $b_{(\pm 2^x)} = 2^{T-1-x}b$, and $b_k = 0$ if $k \not\equiv \pm 2^s \pmod{n}$ for all positive s.

We next consider even n; let $r = n/2$ with $r \geq 2$. The expressions for $\{d_k\}$ become

$$d_{2j-1} = -\sum_{i=1}^{j-1} a_i + \sum_{i=j+1}^{2j-1} a_i - \sum_{i=2j}^{j+r-1} a_i + \sum_{i=j+r+1}^{2r} a_i$$

$$d_{2j} = -\sum_{i=1}^{j} a_i + \sum_{i=j+1}^{2j} a_i - \sum_{i=2j+1}^{j+r} a_i + \sum_{i=j+r+1}^{2r} a_i$$

Equality between successive quantities in $d_1, \ldots, d_{2r-1}$ yields

$$A_j: \quad a_j + a_{r+j} = 2a_{2j} \qquad \text{for } 1 \leq j \leq r-1;$$

$$B_j: \quad a_{j+1} + a_{r+j+1} = 2a_{2j+1} \qquad \text{for } 1 \leq j \leq r-1.$$

Comparing neighboring equations in the sequence $A_1, B_1, A_2, B_2, \ldots, A_{r-1}, B_{r-1}$ yields $b_{r+j} = 2b_{2j}$ for $1 \leq j \leq r-1$ and $b_{2j+1} = 0$ for $1 \leq j \leq r-2$. Together with A_1, which is equivalent to $b_1 = \sum_{i=2}^{r} b_i$, these give a complete set of equations for $\{b_j\}$.

We consider two cases, depending on the parity of r. Suppose first that r is even. When $r = 2$, our equations are $b_1 + b_3 = 2b_2$ and $b_1 = b_2$, which require $b_1 = b_2 = b_3$. Suppose then that $r \geq 4$. In this case $b_{r\pm1} = 0$, and thus $b_1 = 2b_2$ and $b_{2r-1} = 2b_{2r-2}$. For other odd k, we have $b_k = 0$. Expressing A_1 as $2b_2 = \sum_{i=2}^{r} b_i$, we thus obtain $b_2 = \sum_{i=2}^{r/2} b_{2i}$. Letting $b'_j = b_{2j}$ for $1 \leq j \leq r-1$, we obtain $b'_1 = \sum_{i=2}^{r/2} b'_i$, $b'_j + b'_{r/2+j} = b_{2j} + b_{r+2j} = 2b_{4j} = 2b'_{2j}$ for $1 \leq j \leq r/2 - 1$, and $2b'_{2j+1} = 2b_{4j+2} = b_{2j+1} + b_{r+2j+1} = 0$ for $1 \leq j \leq r/2 - 2$. This reduces the case $n = 2r$ with r even to the case where $n = r$. Note that when $n = 4$, all b_k are nonzero, but this fails when $n = 8$ because we obtain $b_3 = b_5 = 0$ before the reduction.

Now suppose that r is odd. Our equations become $b_1 + b_{r+1} = 2b_2$, $b_{r-1} + b_{2r-1} = b_{2r-2}$, $b_1 = \sum_{i=2}^{r-1} b_i$, and $b_k = 2b_{(2k)}$ for even k with $0 < k < 2r$ and $k \neq r \pm 1$, where (u) again denotes the least positive residue of the integer u modulo $n = 2r$. If k is even and $2^s k \not\equiv r \pm 1 \pmod{2r}$ for all s, then $b_k = 2^s b_k$ for some positive s, which yields $b_k = 0$. All other even k with $0 < k < 2r$ satisfy $k \equiv \pm 2^s$ for some $s > 0$. There is a least positive integer T such that $2^T \equiv r \pm 1 \pmod{n}$, and so for $1 \leq s \leq T$ we have $b_{(2^s)} = 2^{T-s}b$ and $b_{(-2^s)} = 2^{T-s}b'$, where $\{b, b'\} = \{b_{r-1}, b_{r+1}$. Now $b_1 + b_{r+1} = 2b_2 = 2^T b$ and $b_1 = \sum_{i=2}^{r-1} b_i = ub + vb'$, where u and v are nonnegative integers with $u > 0$, and $u + v = 2^T - 1$. Hence either $2^T b = ub + (v+1)b'$ or $2^T b = (u+1)b + vb'$. In the former case, $b = b'$. In the latter case, $b = b_{r+1}$; thus the sum for b_1 contains the term $b_{r-1} = b'$, and we obtain $v > 0$ and again $b = b'$.

In all cases we have shown that a nonnegative solution $\{b_k\}$ exists, unique up to multiplication by a constant. The solution contains no zero terms if and only if $n \in \{1, 2, 4\}$ or n is a prime such that -1 and 2 generate the group of nonzero residues modulo n.

Editorial comment. The condition $a_1 \leq a_2 \leq \cdots \leq a_n$ appeared in the problem proposal but was omitted in the published version; all solvers recognized the need for some such condition. The proposers note that in H. Steinhaus, *100 Problems in Elementary Mathematics*, Basic Books, 1964, reprinted by Dover, 1979, Problem 56 calls for minimizing $D(a_1, \ldots, a_k) + D(a_{k+1}, \ldots, a_n)$ for a particular sequence of length 120. The solution in the book says that $k = 60$ minimizes, but $k = 72$ is better (by 0.24). Since any calculus of such problems would include a method of recognizing balanced sequences, it is striking that a deep problem of number theory would appear in the characterization in part (**b**).

Solved also by D. Beckwith, D. E. Knuth, J. H. Lindsey II, and the proposers.

A Sequence of Distinct Integers

10432 [1995, 169]. *Proposed by David M. Bloom, Brooklyn College, CUNY, Brooklyn, NY.* Let

$$P = \{p \in \mathbb{Z}^+ : p \text{ is prime and } p \equiv 3 \pmod 4\}.$$

For $p \in P$, let $S(p)$ denote the sum of all quadratic residues (mod p) that lie in the interval $(0, p/2)$, and let $R(p)$ denote the least positive residue of $S(p) \pmod p$.
(a) Prove that R is one-to-one.
(b) Show that there are infinitely many positive integers that are not in the range of R.

Solution by Thomas Honold, Technical University, Munich, Germany. We have $S(3) = R(3) = 1$. For $p > 3$, we compute $R(p)$ in terms of the residue class of p modulo 16.

$p \pmod{16}$	3	7	11	15
$R(p)$	$\frac{11p-1}{16}$	$\frac{7p-1}{16}$	$\frac{3p-1}{16}$	$\frac{15p-1}{16}$

Let $S'(p)$ denote the sum of all quadratic residues (mod p) that lie in the interval $(p/2, p)$, and let $N(p)$ denote the sum of all quadratic nonresidues in the interval $(0, p/2)$. Since p is congruent to 3 modulo 4 and is not divisible by 3, 24 divides $p^2 - 1$. Also note that $S(p) + S'(p) \equiv \sum_{j=1}^{(p-1)/2} j^2 = p(p^2-1)/24 \equiv 0 \pmod p$.

Since -1 is a quadratic nonresidue (mod p), x is a nonsquare (mod p) if and only if $p-x$ is a square (mod p); thus $S'(p) \equiv -N(p) \pmod p$. Hence $S(p) \equiv N(p) \pmod p$, and we have $2S(p) \equiv S(p) + N(p) = \sum_{k=1}^{(p-1)/2} k = (p^2-1)/8$. We conclude that $16S(p) \equiv -1 \pmod p$. Thus $16R(p) + 1 = kp$, where $1 \le k \le 16$. Since $kp \equiv 1 \pmod{16}$, we have $k = 11, 7, 3, 15$ for $p \equiv 3, 7, 11, 15 \pmod{16}$, respectively.

Suppose that $R(p) = R(q)$ for primes $p, q \in P$. Since $R(p) > 1$ for $p > 3$, we have $p, q > 3$ and $kp = lq \equiv 1 \pmod{16}$ for distinct $k, l \in \{3, 7, 11, 15\}$. Since no prime is a multiple of 5, this is impossible when k or l is 15. Since $\{11, 7, 3\}$ are all prime and $p, q > 3$, the remaining cases are also impossible.

If t is in the range of R, then the prime divisors of $16t + 1$ lie in $P \cup \{5\}$. Since 13 divides $16(13k + 4) + 1$ for each $k \ge 0$, none of the positive integers of the form $13k + 4$ are in the range of R.

Solved also by R. J. Chapman (U. K.), T. R. Hagedorn, R. Holzsager, N. Komanda, J. H. Lindsey II, O. P. Lossers (The Netherlands), L. E. Mattics, P. Venzke, A. N. 't Woord (The Netherlands), NSA Problems Group, and the proposer.

Commutative Algebra without Commutativity

10437 [1995, 170]. *Proposed by J. Maurice Rojas, University of California, Berkeley, CA, and AT&T Bell Laboratories, Naperville, IL.* Let R be a ring (whose multiplication is not necessarily commutative or associative) without zero divisors. Let $x_1, \ldots, x_n$ be algebraically independent indeterminates over R that commute and associate amongst themselves and commute with the elements of R. Also assume the associative law for products of one element of R and two x_i. Prove the following:
(a) If $f \in R[x_1, \ldots, x_n]$ is homogeneous, then any divisor of f is homogeneous.
(b) If $\alpha_1, \ldots, \alpha_n$ are nonzero elements of R and $d_1, \ldots, d_n$ are nonnegative integers with $\gcd(d_1, \ldots, d_n) = 1$, then the polynomial $\alpha_1 x_1^{d_1} + \cdots + \alpha_n x_n^{d_n}$ is irreducible in $R[x_1, \ldots, x_n]$, i. e., every factorization has at most one nonconstant factor.

Solution by the proposer (current affiliation: Massachusetts Institute of Technology, Cambridge, MA). In this solution we assume some familiarity with basic convex geometry. Each nonnegative point $q = (q_1, \ldots, q_n) \in \mathbb{Z}^n$ determines (up to a constant multiple) a monomial $c_q x_1^{q_1} \cdots x_n^{q_n}$, which we write as $c_q x^q$. Thus we may write each $f \in R[x_1, \ldots, x_n]$ as

$f(x) = \sum_{q\in\mathbb{Z}^n} c_q x^q$. Define the *support* supp f of f to be $\{q \in \mathbb{Z}^n : c_q \neq 0\}$. Define the *Newton polytope* $N(f)$ of f to be the convex hull of supp f in $\mathbb{R}^n$. The vertices of a Newton polytope have integral coordinates. Let O be the origin in $\mathbb{R}^n$; observe that $N^{-1}(O) = R$.

We reduce our factorization questions to results on the decomposition of polytopes into vector sums by proving the following lemma.

Lemma. *For any product of polynomials $g_1, \ldots, g_k$, however it is associated, the Newton polytope equals $N(g_1) + \cdots + N(g_k)$.*

Proof. This is immediate for $k = 1$ and the case $k = 2$ is the inductive step in the proof for general k. It remains to give a proof for $k = 2$.

Multiplying monomials in $x_1, \ldots, x_n$ adds exponents, since the variables commute and associate among themselves. Thus supp $g_1g_2 \subseteq$ supp g_1 + supp g_2, and so $N(g_1g_2) \subseteq N(g_1) + N(g_2)$. We need verify only that the coefficients of g_1g_2 corresponding to vertices of $N(g_1) + N(g_2)$ are nonzero.

Let v be such a vertex, and let w be a vector such that v is the only point of $N(g_1)+N(g_2)$ minimizing the inner product $\langle w, v\rangle$. Since v is the vector sum of the faces of $N(g_1)$ and $N(g_2)$ with inner normals equal to w, it is clear that $v = v_1 + v_2$ where v_i is a vertex of $N(g_i)$. Hence the corresponding monomial terms satisfy $c_v x^v = (c_{v_1}x^{v_1})(c_{v_2}x^{v_2})$. Since R has no zero divisors, the coefficient c_v must be nonzero. □

To prove **(a)**, note that f is homogeneous if and only if $N(F)$ is contained in a hyperplane normal to $(1, \ldots, 1) \in \mathbb{R}^n$. Hence every summand must also be contained in a hyperplane normal to $(1, \ldots, 1)$; otherwise the Lemma implies that $N(F)$ would have a face not parallel to this hyperplane.

Let $\mathcal{P}$ be the set of all integral polytopes contained in the closure of the positive orthant of $\mathbb{R}^n$. To prove **(b)**, it suffices by the Lemma to show that $N(\alpha_1x_1^{d_1} + \cdots + \alpha_nx_n^{d_n})$ is indecomposable with respect to vector sum in $\mathcal{P}$, i.e., if $N(\alpha_1x_1^{d_1} + \cdots + \alpha_nx_n^{d_n}) = P_1 + P_2$ for some $P_1, P_2 \in \mathcal{P}$, then either P_1 or P_2 consists only of the origin O. Let s be the number of nonzero exponents d_i, and let $t = \min\{s, n-1\}$. Since $N(\alpha_1x_1^{d_1} + \cdots + \alpha_nx_n^{d_n})$ is the t-simplex whose vertex set is the scaled standard basis $\{d_1\hat{e}_1, \ldots, d_n\hat{e}_n\}$ and $\gcd(d_1, \ldots, d_n) = 1$, it suffices to prove the following theorem.

Theorem. *Suppose a polytope $P \in \mathcal{P}$ satisfies the following three conditions:* (i) *All 2-dimensional faces of P are triangles;* (ii) *$P \cap \{x : x_i = 0\} \neq \emptyset$ for all $i \in \{1, \ldots, n\}$; and* (iii) *γP is not integral for any $\gamma \in (0, 1)$. Then P is indecomposable in $\mathcal{P}$.*

Proof. By Theorem 3 of B. Grünbaum, *Convex Polytopes*, Interscience, 1969, p. 321, (i) implies that every summand of P in a vector sum is a nonzero point or is λP for some $\lambda \in [0, 1]$. By (ii), this implies that a decomposition of P in $\mathcal{P}$ must have the form $\lambda_1 P + \cdots + \lambda_k P$, where $\lambda_i \in (0, 1)$ and $\lambda_1 + \cdots + \lambda_k = 1$. By (iii), all but one of the λ_i are 0. □

Editorial comment. K. S. Kedlaya proved **(a)** by first showing by induction on n that $R[x_1, \ldots, x_n]$ has no zero divisors. This follows, as in section 18 of O. Zariski and P. Samuel, *Commutative Algebra*, Vol. I, Van Nostrand, 1958, by writing polynomials as $\sum a_i x_n^i$ with $a_i \in R[x_1, \ldots, x_{n-1}]$ and considering the terms of highest degree. Then, if $f = g \cdot h$, arrange the terms of the factors by total degree. The product of the parts of smallest degree gives the part of f of smallest degree, and similarly for the parts of largest degree.

REVIEWS

Edited by **Underwood Dudley**

Mathematics Department, De Pauw University, Greencastle, IN 46135

The Sheer Joy of Celestial Mechanics. By Nathaniel Grossman. Birkhäuser, Boston, 1996. 181pp.

Reviewed by J. M. Anthony Danby

Many of us have felt, while teaching calculus, that the "applications" that are embedded into the traditional texts seem sterile, having little relation to the subject matter as it is applied outside the classroom. Professor Grossman has addressed this by assembling some topics from the field of celestial mechanics to show how the techniques of calculus can be "put to work." He also shows how some topics that contemporary students may never encounter (Lagrange expansions, or Bessel functions, for instance) can be used in natural ways to solve problems. The text could be used in an elective course for juniors and seniors. It is clearly a labor of love: hence the title.

The text opens with an introduction to Newtonian mechanics using vectors. There is some emphasis on rotating reference systems, illustrated by a discussion of the Foucault pendulum. Some of this material could be profitably included in the calculus sequence.

The second chapter is concerned with "Central Forces," that is, the motion of a particle subject to a force directed toward a fixed point. This includes discussions of theorems by Bonnet, on the superposition of different force fields, and by Hamilton, on the law of force necessary for a closed elliptic orbit. Bertrand's theorem, proved in a later chapter, is also introduced: this states that the only central laws of force, kr^p, proportional to a power of the distance from the center, for which all bounded orbits are closed, are those with $p = 1$ or -2. This is important in arguing for the uniqueness of the inverse square law of gravitational attraction. It should be noted that although the theorems are stated and proved, the format is conversational and readable. Students allergic to "proofs" will find little to scare them.

Celestial mechanics makes its entry in the third chapter with a discussion of orbits under the inverse square law. Here I felt some frustration. The only equations considered concern attraction to a fixed center, and in this context the author claims to "derive Kepler's Third Law from Newton's Law of Universal Gravitation." Assuming a central force, the equation of motion for mass m_1 around mass m_2 is

$$m_1 \frac{d^2\vec{r}}{dt^2} = -Gm_1m_2\frac{\vec{r}}{r^3}.$$

But the assumption that the central body is unaccelerated violates Newtonian mechanics. If we accept Newton's first and third laws, then a couple of minutes of

derivation produces the correct equation:

$$\frac{d^2\vec{r}}{dt^2} = -G(m_1 + m_2)\frac{\vec{r}}{r^3}.$$

From this is derived the period of an elliptic orbit with semimajor axis a:

$$P = 2\pi\sqrt{\frac{a^3}{G(m_1 + m_2)}}.$$

It is to this formula that astronomers usually refer when they mention "Kepler's third law," and rightly so, since it is responsible for much of our knowledge of stellar masses. I find it a pity that only the pre-Newtonian form of the law should appear in this text. However, the basic variables and equations are clearly introduced and illustrated. The solution of Kepler's equation is made an occasion for discussing the iteration $x_{i+1} = f(x_i)$ for the solution of the equation $x = f(x)$.

An interesting variation appears in a discussion of motion if the constant of gravitation is gradually diminishing. There is some speculation that this is the case; it could have some interesting astrophysical and dynamical consequences. The equation to be solved can be written as

$$\frac{d^2\vec{r}}{dt^2} = \mu(1 - \epsilon t)\frac{\vec{r}}{r^3},$$

where ϵ is *very* small. This is discussed using a perturbation procedure. It may be of interest that if the equation considered is

$$\frac{d^2\vec{r}}{dt^2} = \frac{\mu}{1 + \alpha t}\frac{\vec{r}}{r^3}, \text{ i.e., } \frac{d^2x}{dt^2} = -\frac{\mu}{1 + \alpha t}\frac{x}{r^3}, \frac{d^2y}{dt^2} = -\frac{\mu}{1 + \alpha t}\frac{y}{r^3},$$

$$\text{where } r = \sqrt{x^2 + y^2},$$

then transformations, due to Meshchersky, to new coordinates, ξ, η and a new "time," τ:

$$\xi = \frac{x}{1 + \alpha t}, \quad \eta = \frac{y}{1 + \alpha t}, \quad \tau = \frac{1}{\alpha(1 + \alpha t)},$$

produce

$$\frac{d^2\xi}{d\tau^2} = -\mu\frac{\xi}{\rho^3}, \quad \frac{d^2\eta}{d\tau^2} = -\mu\frac{\eta}{\rho^3}, \quad \text{where } \rho = \sqrt{\xi^2 + \eta^2}.$$

Thus, the original problem can be discussed completely using the known solutions of the transformed equations.

Chapter four, on expansions in elliptic motion, demonstrates that there is more to expansions than Taylor's theorem! Lagrange's expansion theorem, Bessel functions, Fourier series, and complex variables are all introduced and used. Some of the algebra is none too friendly, but the methods used are not advanced and the text is readable.

Chapter five includes a discussion of Bertrand's theorem and some application of differential geometry in the plane.

The remainder of the text deals with solid bodies and potential theory. We move from two dimensions to three. Chapter six includes a discussion of moments of inertia, Eulerian angles, and Euler's equations, with applications to the consequences of the non-sphericity of the Earth on its rotation. Chapter seven is not so easy to follow, but is needed in part for later applications. It is concerned with gravitational potential of ellipsoids. Applications include a discussion of the precession of the equinoxes and the nutation. Finally, in chapter eight there are discussions of the tides, variation of gravity on the surface of the Earth, and ellipsoidal figures of rotating fluid bodies. Over the last century there has been beautiful work in this area by the likes of Poincaré, Darwin, Jeans, and Chandrasekhar; it is good to see some of this discussed here.

The text includes many problems. Certainly, to start with, students will find these difficult, since they call for understanding and applying the mathematics, not just using formulas. Unfortunately in the latter part of the text, the number of problems falls off, due, I must add, to the nature of the material and not to any failure on the part of the author.

In my opinion, the text does not constitute a good introduction to the subject matter of celestial mechanics. The author has made his own selection based on his personal taste; but too much standard elementary material has been omitted. Certainly the text would be unsuitable for any student concerned with gaining practical ability in celestial mechanics. For a student who might be interested in working on the mathematical reaches of the subject, I would recommend the text *Mathematical Introduction to Celestial Mechanics* by Harry Pollard, Mathematical Association of America, 1976.

The problems are, indeed, old-fashioned, and there is a lot of nostalgia about this text. I think that the author's rationale for writing it and using it are valid, and recommend its consideration for other potential instructors. But any instructor must know the material well and be committed to transmitting it as something to be enjoyed. It will be a personal matter; but have a look at the text and give it a chance.

Department of Mathematics
North Carolina State University
Raleigh, NC 27695
danby@math.ncsu.edu

Strength in Numbers: Discovering the Joy and Power of Mathematics in Everyday Life. By Sherman K. Stein. John Wiley & Sons, 1996, xiii + 272, $24.95.

Reviewed by **Jennifer R. Galovich**

On a recent visit to the local McBookstore, and in a somewhat curmudgeonly mood, I decided to check out the range of mathematics and science offerings. I expected to come away with even more evidence of the general hopelessness of engaging the public imagination with respect to mathematics, sure that the shelves would be loaded with Stephen Jay Gould's excellent books and that there would be little room left over for the "hard" sciences.

Here are the results of my completely unscientific survey:

Category	Number of titles
Life Science	92
Mathematics	85
Physics	44
Geology	19
Chemistry	3

Well! We don't do so badly after all! I was pleasantly surprised to find that while Professor Gould was indeed well represented, many mathematics titles were also stocked. Moreover, there were quite a few more mathematics books classified as General Science, apparently depending on whether or not the word "mathematics" actually appeared in the title. While the volume under review was not yet on the shelves, it deserves to be. This book has something for almost everybody.

Strength in Numbers is divided into three parts, the first of which is a compendium of essays for a very general audience. Two chapters on the seductive power of numbers are followed by a chapter on how, nevertheless, numbers can be an antidote to clever rhetoric. Stein also includes an unusual section on debunking myths—some quasi-historical ("As a boy, Einstein was poor in arithmetic." [p. 42]), some cultural ("Mathematics is a dead subject." [p. 33]). Mathematicians who have heard such stories for years might be interested to see if they agree with Stein's arguments.

Educators in general, and high school teachers in particular, will find good uses for the extensive table describing the mathematical skills required for various occupations [pp. 65–68]. However the sections of most interest to educators and other readers of this journal are those in which Stein reviews the history of reform in mathematics education in this century. He pays special attention to the SMSG curriculum of the 1960s (we know how *that* came out) and the more recently developed NCTM Standards. Stein finds these standards ambitious—but are they more ambitious than they need to be? Stein's views are controversial; given his longstanding interest and experience in education at all levels, they are also worth reading.

In the second part, Professor Stein takes a fresh look at some topics accessible to anyone with two years of high school mathematics (Algebra I and Geometry). Beginning with the definition of a prime number, Stein offers the stories of the Mertens and Pólya Conjectures as an object lesson in why it's not enough just to check a few (or even a large number of) cases. A tutorial in which the reader "discovers" the formula for the sum of a geometric series is paired with a discussion of the money multiplier in economics. And a quick review of the arithmetic of fractions is followed by discussions of irrational numbers (including Euclid's proof that $\sqrt{2}$ is irrational) and Cantor's diagonalization argument.

The final part makes more demands on the reader's staying power. It consists of a quick tour through the derivative ("How Steep is a Curve?") and the integral ("Finding a Curved Area"), which relies heavily on some earlier chapters. The section culminates in a proof due to Hindu mathematicians of the sum of Gregory's series:

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots.$$

The proof is pretty strenuous for a general reader, but would be a good project for calculus students.

Professor Stein is well-known as an excellent expositor. He has a gift for explaining a difficult concept by finding an image that carries the central idea. For example, his description of the Radon transform as a theoretical way of locating all the different fruits and nuts in a fruitcake without cutting it open is particularly nice. (Some readers may prefer that the fruitcake simply be transformed).

There are a few places, however, where level could be a problem. For example, in the chapter entitled "The Mother of Invention" we move from the definition of whole numbers to the statement of Euler's generalization of Fermat's Little Theorem in less than a page. The required learning curve here is rather steep!

In writing for a general audience it can be difficult to decide how much subtlety should be mentioned and how much should be airbrushed. In general, Stein handles such dilemmas well. The sum of a geometric series, for example, makes an early appearance and is used in several later chapters. How much should one say about what it means to "add up" infinitely many numbers? Stein disposes of this in one neat sentence, conveying the key idea without getting sidetracked.

Purists may balk at Stein's take on the "Let's Make a Deal" problem. Although he steps over some of the subtleties (see, e.g., **[1]**), I found the presentation very effective for the intended audience. He recounts the original problem but not the answer, and promises readers that they can find the solution and explain it—they just need an opportunity and some guidance in thinking mathematically. Stein provides both by giving clear directions on how to simulate the problem and suggesting questions one might ask along the way.

Sherman Stein has produced an eminently readable book that should appeal to a wide audience. The curmudgeonly view is that while *Strength in Numbers* is eminently readable, it doesn't follow that it will actually be read. After all, while book sales are up, we suspect that book reading may be down. On the other hand, if we despair at the appearance of yet another eminently readable book and wonder who will read it, then perhaps we might think in a more focused way about how to change that state of affairs. Susan Landau's editorial in the *Notices* of the American Mathematical Society **[2]** calls us to "rise to the challenge" implicit in the public's increasing awareness of and interest in mathematics. Indirectly, she reminds us of our collective and individual responsibility as professionals to profess. Some of us profess in the classroom, others as writers, and most of us do some of both. More and more of us are involved in activities that help the public understand why mathematics is useful and exciting. So I applaud this new book; its appearance reminds us—in more than one way—that there really can be "strength in numbers."

REFERENCES

1. L. Gillman, The car and the goats, this MONTHLY **99** (1992) 3–7.
2. S. Landau, *Notices Amer. Math. Soc.* **43** (1996) 652.

St. John's University
Collegeville, MN 56321
jgalovich@csbsju.edu

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**

with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1–4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, S(14–16), L*. *101 Careers in Mathematics.* Ed: Andrew Sterrett. MAA, 1996, x + 260 pp, $20 (P). [ISBN 0-88385-704-9] A selection of informal, personal profiles reprinted from the "mathematician of the month" series that editor Sterrett produced for several years: ordinary math majors who work in a variety of careers, generally requiring only bachelor's or master's degrees. A terrific resource for undergraduates who want to know what they might do with a math major. Concludes with job-seeking advice reprinted from *Math Horizons.* LAS

Finite Mathematics, T(13–14: 1). *Finite Mathematics, Models, and Structure.* William J. Adams. Kendall/Hunt, 1995, xii + 437 pp, $52.44 (P). [ISBN 0-7872-0995-3] Chatty, nicely written text. Topics: modeling, matrices, linear programming, probability and Bernoulli trials, introduction to game theory, logical deduction, etc. RM

Education, P, L*. *The Nature of Mathematical Thinking.* Eds: Robert J. Sternberg, Talia Ben-Zeev. Stud. in Math. Thinking & Learning Ser. Lawrence Erlbaum Associates, 1996, xiv + 335 pp, $34.50 (P); $69.95. [ISBN 0-8058-1799-9; 0-8058-1798-0] Why do some find mathematics easy, others painful? Why are some good at algebra, but terrible with geometry? Why are some successful at business, but no good in mathematics? These eleven independently authored chapters attempt but fail to reach consensus in response to these questions. Their diverse perspectives—psychometric, cognitive, educational, cultural, and mathematical—have virtually nothing in common. Yet as a whole they offer a wealth of hypotheses and insights. LAS

History, P, L*. *Modern Mathematics in the Light of the Fields Medals.* Michael Monastyrsky. AK Peters, 1997, xv + 160 pp, $35. [ISBN 1-56881-065-2] Expanded English translation of a remarkable paper first published in Russia in 1991 that provides concise (high-level) expository surveys of the mathematical work of every Fields medalist since the prize was first awarded in 1936. Includes historical background on Fields, on the prize, and on the selection process. Two appendices bring the work up-to-date. LAS

Logic, P, L. *Deviant Logic, Fuzzy Logic: Beyond the Formalism.* Susan Haack. Univ of Chicago Pr, 1996, xxvi + 291 pp, $18.95 (P); $55. [ISBN 0-226-31134-1; 0-226-31133-3] Revision of 1974 Cambridge University Press edition (TR, June–July 1975). Adds 5 papers on related topics and updates bibliography; remaining material essentially unchanged. Argues that while classical logic may need revision, none of the proposed alternatives (e.g., quantum logics, modal logics, fuzzy logic) is satisfactory. Primarily philosophical, rather than mathematical, treatment. LB

Logic, P, L. *The Principles of Mathematics Revisited.* Jaakko Hintikka. Cambridge Univ Pr, 1996, xii + 288 pp, $59.95. [ISBN 0-521-49692-6] Philosophical essay challenging basic assumptions underlying standard first-order logic. Explores consequences of a new first-order logic (modifies handling of scope of quantifiers) for foundations of mathematics. Well-

written and accessible to undergraduates with some background in logic. LB

Logic, P. *Sentential Probability Logic: Origins, Development, Current Status, and Technical Applications.* Theodore Hailperin. Lehigh Univ Pr, 1996, 304 pp, $43.50. [ISBN 0-934223-45-9] Studies a logic in which probabilities play a semantic role comparable to truth values in conventional logic. Includes a thorough historical survey and applications. LB

Combinatorics, P. *Codes, Designs and Geometry.* Ed: Vladimir Tonchev. Kluwer Academic, 1996, 120 pp, $115. [ISBN 0-7923-9759-2] Papers from a workshop held in 1994 at Michigan Technological University.

Discrete Mathematics, P, L*.** *Spectral Graph Theory.* Fran R.K. Chung. CBMS Reg. Conf. Ser. in Math., No. 92. AMS, 1997, xi + 207 pp, $25 (P). [ISBN 0-8218-0315-8] The spectrum of a graph is the set of eigenvalues of a particular matrix associated with the graph. This monograph presents the use of the graph spectrum to investigate properties and structures of a graph. Emphasizes important connections to geometry, but requires no special background in geometry. Beautifully clear exposition and careful attention to applications and to connections with other areas of mathematics. LB

Number Theory, T(14: 1), S*, L. *A Pathway Into Number Theory, Second Edition.* R.P. Burn. Cambridge Univ Pr, 1997, xv + 262 pp, $27.95 (P). [ISBN 0-521-57540-0] Introduction to number theory through graded problems. New material on RSA codes, Gaussian integers, and triangular numbers. Expanded historical notes. (*First Edition*, TR, December 1982.) DB

Linear Algebra, T(17: 2), L. *Linear Algebra.* Peter D. Lax. Pure & Appl. Math. Wiley, 1997, xiv + 250 pp, $54.95. [ISBN 0-471-11111-2] Graduate-level introduction to linear algebra. Standard topics as well as chapters on matrix inequalities, kinematics and dynamics, convexity, the duality theorem, and positive matrices. LC

Linear Algebra, T(14: 1).** *Linear Algebra with Applications.* Otto Bretscher. Prentice Hall, 1997, xiii + 587 pp. [ISBN 0-13-190729-8] Appealing order of topics: linear systems and transformations, subspaces of $\mathbb{R}^n$, orthogonality, determinants, eigenvectors, coordinate systems, differential equations, linear spaces. Uses dynamical systems as unifying theme. Full of nice applications. Theory motivated by examples. Great problems. TH

Group Theory, S(14–15). *Laboratory Experiences in Group Theory.* Ellen Maycock Parker. MAA, 1996, xi + 81 pp, $22 (P), with disk. [ISBN 0-88385-705-7] 15 labs that encourage students to make conjectures and to become comfortable with open-ended questions. "Exploring Small Groups" software is on the accompanying disk. LC

Group Theory, T(18: 1), P, L. *Lectures on Exceptional Lie Groups.* J.F. Adams. Ed: Zafer Mahmud, Mamoru Mimura. Lect. in Math. Ser. Univ of Chicago Pr, 1996, xiv + 122 pp, $25 (P). [ISBN 0-226-00527-5] Detailed lectures given at Cambridge on the construction of exceptional Lie groups, their representations, and their interconnections. A sequel to *Lectures on Lie Groups*, J.F. Adams (1982). TH

Algebra, T(14), L. *Introductory Modern Algebra: A Historical Approach.* Saul Stahl. Wiley, 1997, xii + 322 pp, $62.95. [ISBN 0-471-16288-4] Introduction to algebra with topics chosen according to historical development, focusing on the problem of solvability by radicals. Chapters 1–3 set up solvability culminating in a constructibility proof of Gauss; 4–7 prove Galois' primitive element theorem; 8–11 concern group theory, especially relating permutation groups to solvability. Exercises; some with solutions. JD

Algebra, T(18: 2), P. *Integer-Valued Polynomials.* Paul-Jean Cahen, Jean-Luc Chabert. Math. Surv. & Mono., V. 48. AMS, 1997, xix + 322 pp, $75. [ISBN 0-8218-0388-3] If the integral domain D has quotient field K, this text studies the ring of polynomials with coefficients in K that map a subset of K into D. LC

Algebra, P. *Prüfer Domains.* Marco Fontana, James A. Huckaba, Ira J. Papick. Pure & Appl. Math., V. 203. Marcel Dekker, 1997, ix + 328 pp, $150. [ISBN 0-8247-9816-3] Comprehensive study of Prüfer domains, including overrings, Dedekind domains, trace properties, and generalizations of Prüfer domains. TH

Algebra, P. *Zariskian Filtrations.* Li Huishi, Freddy van Oystaeyen. K–Mono. in Math., V. 2. Kluwer Academic, 1996, ix + 252 pp, $127. [ISBN 0-7923-4184-8]

Algebra, T(15–16: 1), L. *Rings, Fields, and Vector Spaces: An Introduction to Abstract Algebra via Geometric Constructibility.* B.A. Sethuraman. Undergrad. Texts in Math. Springer-Verlag, 1997, xiii + 190 pp, $34.95. [ISBN 0-387-94848-1] Uses geometric constructibility as motivation. Intended as a one-semester abstract algebra course for future teachers. No group theory. Worth a look. LC

Algebra, P. *Commutative Ring Theory.* Eds: Paul-Jean Cahen, *et al.* Lect. Notes in Pure & Appl. Math., V. 185. Marcel Dekker, 1997,

xii + 470 pp, $175 (P). [ISBN 0-8247-9815-5] Proceedings of the Second International Conference on Commutative Ring Theory held in Fès, Morocco.

Algebra, P. *Operads: Proceedings of Renaissance Conferences.* Eds: Jean-Louis Loday, James D. Stasheff, Alexander A. Voronov. Contemp. Math., V. 202. AMS, 1997, ix + 443 pp, $85 (P). [ISBN 0-8218-0513-4] Papers from two 1995 events: a special session in Hartford, Connecticut, and a conference in Luminy, France.

Calculus, T(13–14). *Fundamentals of Calculus with Applications.* William J. Adams. Kendall/Hunt, 1996, xiv + 391 pp, (P). [ISBN 0-7872-1115-X] A standard introduction to calculus for students in business, *et al.* Well-written; excellent examples and clever illustrations. Worth considering. PF

Complex Analysis, P. *Extremal Riemann Surfaces.* Eds: J.R. Quine, Peter Sarnak. Contemp. Math., V. 201. AMS, 1997, xz + 243 pp, $49 (P). [ISBN 0-8218-0514-2] Papers solicited for an AMS Special Session at the 1995 meeting in San Francisco.

Differential Equations, T(15), L. *Computational Differential Equations.* K. Eriksson, *et al.* Cambridge Univ Pr, 1996, xvi + 538 pp, $100; $44.95 (P). [ISBN 0-521-56312-7; 0-521-56738-6] Finite element methods for linear ordinary and partial differential equations. Combines mathematical analysis and computation. Develops topics of error estimation and adaptive error control throughout. AO

Numerical Analysis, T(15: 2), L. *Theory and Applications of Numerical Analysis, Second Edition.* G.M. Phillips, P.J. Taylor. Academic Pr, 1996, xii + 447 pp, $24.95 (P). [ISBN 0-12-553560-0] Two new chapters: splines and other approximations; matrix eigenvalues and eigenvectors. Computing exercises have been added at the end of each chapter. (*First Edition,* TR, November 1974.) AO

Numerical Analysis, P. *Level Set Methods.* J.A. Sethian. Cambridge Univ Pr, 1996, xviii + 218 pp, $39.95. [ISBN 0-521-57202-9] State-of-the-art computational techniques for modeling the evolution of boundaries and interfaces in a variety of application areas (e.g., burning flames, ocean waves, medical imaging, grid generation). AO

Functional Analysis, P. *Functional Analysis.* Eds: Susanne Dierolf, Seán Dineen, Paweł Domański. Walter de Gruyter, 1996, xi + 473 pp, DM 268. [ISBN 3-11-014617-7] Proceedings of a 1994 workshop at Trier University, Germany.

Algebraic Geometry, P. *Collected Papers of Giacomo Albanese.* Eds: Ciro Ciliberto, Paulo Ribenboim, Edoardo Sernesi. Papers in Pure & Appl. Math., V. 103. Queen's Univ, 1996, xii + 182 pp, (P). [ISBN 0-88911-737-3]

Differential Geometry, P. *Sub-Riemannian Geometry.* Eds: André Bellaîche, Jean-Jacques Risler. Prog. in Math., V. 144. Birkhäuser Boston, 1996, viii + 393 pp, $84.50. [ISBN 0-8176-5476-3] 5 articles provide an introduction to the field and survey the state-of-the-art.

Algebraic Topology, T(18), P. *Sheaf Theory, Second Edition.* Glen E. Bredon. Grad. Texts in Math., V. 170. Springer-Verlag, 1997, xi + 502 pp, $59.95. [ISBN 0-387-94905-4] Sheaf-theoretic cohomology from an algebraic topology point-of-view. Assumes substantial background in homological algebra and algebraic topology. Several cohomology theories are compared, applications of spectral sequences given. New edition includes Oliver transfer, Conner conjecture, intersection theory. Exercises; some with solutions. JD

Game Theory, P, L*. *Games of No Chance.* Ed: Richard J. Nowakowski. Math. Sci. Res. Inst. Public., V. 29. Cambridge Univ Pr, 1996, xiii + 537 pp, $49.95. [ISBN 0-521-57411-0] Papers from a 1994 workshop on combinatorial game theory at MSRI in Berkeley. Includes introductory papers, papers on "classical" games (e.g., chess and Go), as well as others. Concludes with a list of open problems and a comprehensive bibliography.

Optimal Control, T(16–17). *Optimal Control: Basics and Beyond.* Peter Whittle. Ser. in Systems & Optimiz. Wiley, 1996, ix + 464 pp, $49.95 (P). [ISBN 0-471-95679-1] "Basics" includes classical topics (e.g., stability, feedback, controllability) as well as more general optimization techniques (e.g., dynamic programming, the Pontryagin maximum principle). "Beyond" topics: risk-sensitive and H_∞ criteria; time-integral methods and optimal stationary policies; near-determinism and large deviation theory. AO

Probability, S(13–15). *Your Intuition Is Wrong!* Marc T. Simon. Dorrance Pub, 1996, xi + 47 pp, $9 (P). [ISBN 0-8059-3834-6] Two dozen wordy word problems in elementary probability embeded in gaming and casino contexts. Mostly standard types: birthday problems, stopping strategies, runs, coin tossing, urns, etc. Concludes with brief explanations at the end. LAS

Mathematical Statistics, T(17–18: 1). *Aspects of Statistical Inference.* A.H. Welsh. Ser. in Prob. & Stat. Wiley, 1996, xviii + 451 pp, $59.95. [ISBN 0-471-11591-6] Includes several non-standard topics: robustness, randomization, finite population inference, computational methods based on simulation and smoothing methods. Simple data sets used to motivate topics and emphasize practical aspects of inference. Assumes some distribution theory. RS

Statistical Methods, S(17–18), P. *Robust Statistical Procedures, Second Edition.* Peter J. Huber. CBMS–NSF Reg. Conf. Ser. in Appl. Math., V. 68. SIAM, 1996, ix + 67 pp, $18.50 (P). [ISBN 0-89871-379-X] Brief, well-organized introduction and overview of robust statistics (*First Edition*, TR, April 1978). This edition adds a chapter on recent developments and updates the list of references. RS

Statistical Methods, T(16–17: 1, 2). *Analysis of Variance, Design and Regression: Applied Statistical Methods.* Ronald Christensen. Chapman & Hall, 1996, xvi + 587 pp, $54.95. [ISBN 0-412-06291-7] Includes matrix formulation of regression and ANOVA models. Examples used to motivate theory rather than simply to illustrate techniques. Minitab output and commands incorporated throughout. Readable. Accessible to students with weaker mathematical background. RS

Statistical Methods, S(18), P. *Modern Multidimensional Scaling: Theory and Applications.* Ingwer Borg, Patrick Groenen. Ser. in Stat. Springer-Verlag, 1997, xvii + 471 pp, $54.95. [ISBN 0-387-94845-7] Comprehensive presentation of multidimensional scaling (MDS). Five parts: (1) introduction—useful for individuals with applied interests, (2) technical aspects of MDS, (3) unfolding as a special case of MDS, (4) geometry of MDS, (5) techniques/models closely associated with MDS. Appendix describes major software packages. KB

Algorithms, P. *Proceedings of the Eighth Annual ACM–SIAM Symposium on Discrete Algorithms.* ACM & SIAM, 1997, x + 788 pp, $79.50 (P). [ISBN 0-89871-390-0]

Theory of Computation, T?(16–17: 1), S, P. *Communication Complexity.* Eyal Kushilevitz, Noam Nisan. Cambridge Univ Pr, 1997, xiii + 189 pp, $37.95. [ISBN 0-521-56067-5] Summary of the mathematical development (with nice practical examples and motivation) of the theory of what needs to be communicated and the amount of communication necessary to solve a problem. Begins with Yao's two-party model, extended to more general models. RM

Computer Science, P, L. *Handbook of Applied Cryptography.* Alfred J. Menezes, Paul C. van Oorschot, Scott A. Vanstone. Ser. on Disc. Math. & Its Applic. CRC Pr, 1997, xxviii + 780 pp, $79.95. [ISBN 0-8493-8523-7] An encyclopedia of practical cryptologic techniques (conventional and public-key). Detailed descriptions of algorithms and extensive references to the research literature. AO

Computer Science, P*, L.** *Selected Papers on Computer Science.* Donald E. Knuth. CSLI Lect. Notes, No. 59. Center for the Study of Language & Information (Leland Stanford Junior Univ., Stanford, CA 94305) & Cambridge Univ Pr, 1996, xii + 274 pp, $69.95; $24.95 (P). [ISBN 1-881526-92-5; 1-881526-91-7] From the Preface: "This book assembles under one roof all of the things I've written about computer science for people who aren't necessarily specialists in the subject."

Applications (Fluid Mechanics), P. *Homogenization and Porous Media.* Ed: Ulrich Hornung. Interdisc. Appl. Math., V. 6. Springer-Verlag, 1997, xvi + 279 pp, $59.95. [ISBN 0-387-94786-8] 10 articles survey the method of homogenization applied to problems in porous media.

Applications (Physical Science), P. *Inverse Problems in Geophysical Applications.* Eds: Heinz W. Engl, Alfred K. Louis, William Rundell. SIAM, 1997, x + 303 pp, $81 (P). [ISBN 0-89871-381-1] Proceedings of a 1995 GAMM–SIAM conference in Yosemite, California.

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–222: Control Using Logic-Based Switching.* Ed: A. Stephen Morse. Springer-Verlag, 1997, viii + 276 pp, $54 (P). [ISBN 3-540-76097-0] 23 papers from a 1995 workshop held on Block Island, Rhode Island.

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–208: Proceedings of Workshop on Advances in Control and its Applications.* Eds: H.K. Khalil, J.H. Chow, P.A. Ioannou. Springer-Verlag, 1996, xxiii + 319 pp, $59 (P). [ISBN 3-540-19993-4] Papers from a 1994 event at the University of Illinois, Urbana–Champaign.

Reviewers

KB: Karla Ballman, Macalester; LB: Lynne Baur, Carleton; DB: David Bressoud, Macalester; LC: Laura Chihara, St. Olaf; JD: Jill Dietz, St. Olaf; PF: Paul Froeschl, Macalester; TH: Tom Halverson, Macalester; RM: Richard Molnar, Macalester; AO: Arnold Ostebee, St. Olaf; RS: Richard Single, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf.

THE AUTHORS

S. C. COUTINHO was born in Recife, Brazil, and received his undergraduate education from the local Federal University. He obtained his Ph.D. from the University of Leeds, U.K., with a thesis supervised by J. T. Stafford. He now lectures at the Department of Pure Mathematics of the Federal University of Rio de Janeiro. His research interests include algebraic K-theory, non-commutative algebra, and the theory of $\mathscr{D}$-modules. He is an omnivorous reader, with a passion for English prose and poetry, past and present.

TIBERIU TRIF received his M.S. degree in mathematics in 1995 at the Babeş-Bolyai University in Cluj-Napoca, Romania. He is now an Assistant at the same university and has started to work on a Ph.D. thesis in theoretical functional analysis under the direction of Professor Wolfgang W. Breckner.

JOHN CHOLLET obtained his B.A. from U.C. Berkeley, his M.A. at San Francisco State University, and his Ph.D. from U.C. Santa Barbara under the direction of Marvin Marcus. He held a post-doctoral fellowship at the University of British Columbia, where he remained as a lecturer before taking a permanent position at Towson State University. His mathematical interests are in multilinear algebra and matrix inequalities. In 1982, he published in the MONTHLY a still unresolved conjecture on a matrix inequality.

EUGENE GUTKIN was educated in Moscow and received his Ph.D. in Boston, from Brandeis University. He has taught in Salt Lake City, New York (Columbia University), Berkeley, and Los Angeles (University of Southern California). He has published research articles on Lie groups, quantum mechanics, and dynamical systems. His current field of mathematical activity is billiard dynamics. Besides mathematics, he derives enjoyment from intermittent juggling and jogging.

STEVEN H. WEINTRAUB received his Ph.D. from Princeton University in 1974 and has been at Louisiana State University ever since, with leaves to visit UCLA, Rutgers, Oxford, Yale, Göttingen, Bayreuth, and Hannover. He has authored or co-authored over 40 research papers in the fields of differential topology and algebraic geometry, and three books—a graduate algebra text, a research monograph on Siegel modular varieties, and an undergraduate text on differential forms that he modestly hopes will revolutionize the teaching of multivariable calculus. He has served on the Council of the AMS and is now chair of the AMS Working Group on Public Awareness of Mathematics. He is an instrument-rated private pilot, and looks forward to purchasing an airplane when and if he wins the lottery. A priori, this event has a low probability; since he doesn't play, the probability is even lower.

KEITH DEVLIN is Dean of the School of Science at Saint Mary's College of California, in Moraga, California, a Senior Researcher at the Center for the Study of Language and Information at Stanford University, and a Consulting Research Professor in the Department of Information Science at the University of Pittsburgh, a range of institutions that reflect his interests in mathematics, logic, applications of mathematics to language and communication, and mathematics and science education. A native of England, he obtained his Ph.D. in mathematical logic at the University of Bristol in 1971. Since 1987, he has resided in the United

States. He is the current editor of the MAA newsletter FOCUS and the author of *Devlin's Angle*, a monthly column on *MAA Online*.

KENT G. MERRYFIELD received his B.A. in Chemistry and Mathematics from Rice University in 1975 and his Ph.D. from the University of Chicago in 1980, under the direction of Alberto P. Calderón. While a student at Rice and Chicago, he managed to take classes from both Salomon Bochner and Antoni Zygmund. He has been a faculty member at California State University, Long Beach since 1985.

VIET NGO was born in 1956 in Hue, Vietnam. He received a B.S. in Mathematics from the University of Minnesota in 1976 and a Ph.D. from the University of California at Berkeley in 1984. He is Associate Professor of Mathematics at California State University, Long Beach. His interests include functional analysis, parallel processing, and Buddhism.

SALEEM WATSON was born in 1952 in Zaytoun, Egypt. He received a B.S. in Mathematics in 1972 from Andrews University in Michigan and a Ph.D. in 1978 from McMaster University in Ontario. He subsequently did research and taught at the Mathematics Institute of Warsaw University in Poland and the Pennsylvania State University. He has been a faculty member at California State University, Long Beach since 1986. His research interests are in topology and functional analysis.

JOSÉ L. PALACIOS is professor in Universidad Simón Bolívar, where he got his B.S. in Mathematics in 1976 and has been teaching ever since except for five years in Berkeley to get his Ph.D. under the supervision of David Aldous (1982) and five years of toiling in the New Jersey Institute of Technology. He is now trying to publish his fourth book of short fiction and is involved with random walks on graphs.

HENRYK GZYL completed a degree in physics at the Universidad Central de Venezuala and a degree in mathematics at UCSD under the influence of the early seventies and R.K. Getoor.

TONY DANBY received degrees in mathematics from Oxford and Manchester University. He was the first chair oboist in the London Philharmonic Orchestra before coming to the U.S. He taught in the astronomy departments at the University of Minnesota and Yale before joining the department of mathematics at N.C. State. He is the author of texts and software in the areas of dynamical astronomy and numerical applications of differential equations. He was recently a winner in the "Computers in Physics" annual Educational Software Contest. In 1977, Minor Planet Danby was named in his honor.

JENNIFER GALOVICH received her undergraduate degree from Reed College in 1969, Master's degree from Brown University in 1972 and, with time off for good behavior, her Ph.D. from the University of Minnesota in 1993. Her research activities are in algebraic and enumerative combinatorics, with a special interest in permutation statistics. She is now on the faculty of St. John's University in Collegeville, MN, where her non-mathematical interests include classical music, detective fiction, and staying warm.

LESTER R. FORD AWARDS FOR 1996

The Lester R. Ford Awards, established in 1964, are made annually to authors of outstanding expository papers in the MONTHLY. The awards are named for Lester R. Ford, Sr., a distinguished mathematician, editor of the MONTHLY (1942–46), and President of the Mathematical Association of America (1947–48).

Winners of the Lester R. Ford awards for expository papers appearing in Volume 103 (1996) of the MONTHLY are:

Robert G. Bartle, Eastern Michigan University.
Return to the Riemann Integral, pp. 625–632.

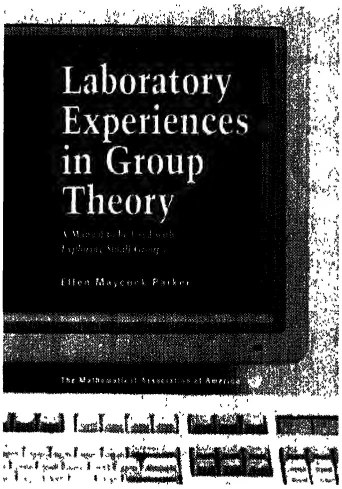
Every calculus student sees the Riemann integral, but it is not flexible or general enough for more technical problems: not enough functions are Riemann integrable, and interchanging limits with a Riemann integral is difficult in the absence of uniform convergence. The Lebesgue integral handles these problems, but it requires the separate study of measure theory and still fails to include some important improper integrals. In this paper, Bartle shows that a generalized Riemann integral captures the advantages of both the Riemann integral and the Lebesgue integral without incurring the major disadvantages of either of the two classical approaches. Bartle points out that the generalized Riemann integral is more general than the Lebesgue integral, in the sense that the set of Lebesgue integrable functions is strictly contained in the set of generalized Riemann integrable functions. The author derives a strong form of the fundamental theorem of calculus and shows how measure theory may be recovered from the theory of generalized Riemann integrals. He demonstrates the advantages of the generalized Riemann integral in dealing with improper integrals and convergence theorems. By the end of the paper, Bartle has presented a strong argument for replacing the Lebesgue integral with the generalized Riemann integral.

A. F. Beardon, University of Cambridge.
Sums of Powers of Integers, pp. 201–213.

The square of the sum of the numbers from 1 *to the given number equals the sum of the cubes of the numbers from* 1 *to the given number*: that was how Levi ben Gerson, in the 14th century, stated the identity that is the starting point for this paper. If we write $\sigma_k(n)$ for the sum of the k-th powers from 1 to n^k, this remarkable identity translates to $\sigma_3(n) = \sigma_1(n)^2$. It cries out for generalization, and Beardon's goal in this paper is to find *all* polynomial relations between σ_i and σ_j, for all i and j. He achieves this by interpreting the question in terms of the elementary theory of algebraic curves: a polynomial relation $T(\sigma_i, \sigma_j) = 0$ means that the curve defined by $T(x, y) = 0$ contains all the points $(\sigma_i(n), \sigma_j(n))$. On the way to the main theorem, we meet Bernoulli numbers, Faulhaber polynomials, and several basic results about algebraic curves. By immersing his question in a general theory, Beardon not only finds a complete answer, but also shows us the power and the charm of the theory of algebraic curves.

John Brillhart, University of Arizona, and **Patrick Morton**, Wellesley College.
A Case Study in Mathematical Research: The Golay-Rudin-Shapiro Sequence, pp. 854–869.

There's a special feel to being on the trail of a discovery, and this paper conveys that feeling to its readers. The basic topic is just a simple sequence of 1's and -1's, and the first observation is that in the first 32,000 steps there are always more 1's than -1's. Yet there is no discernible reason why this should be true. To understand it, Brillhart and Morton have to look deeper, experimentally detecting a "wave-like" pattern in the excesses and describing it precisely enough that its existence can be proved in general. But this is just the beginning. The excesses $s(n)$ are observed to be about the same size as $\sqrt{n}$. Computation with special cases shows that the ratio $s(n)/\sqrt{n}$ can approach any value between $\sqrt{3/5}$ and $\sqrt{6}$; does it always stay in that small range? They piece together an argument for the upper bound and modify it to get the lower bound—until at the last step the method absolutely refuses to work. Undeterred, they look around for a new approach, finally discovering several "mysterious and elegant properties" that produce the final proof. Students who want to know what mathematical research feels like can find out here.



THE MATHEMATICAL ASSOCIATION OF AMERICA

Laboratory Experiences in Group Theory

A Manual to be Used with *Exploring Small Groups*

Ellen Maycock Parker

Series: Classroom Resource Materials

A lab manual with software for introductory courses in group theory or abstract algebra

Laboratory Experiences in Group Theory is a workbook of 15 laboratories designed to be used with the software *Exploring Small Groups* as a supplement to the regular textbook in an introductory course in group theory or abstract algebra. Written in a step-by-step manner, the laboratories encourage students to discover the basic concepts of group theory and to make conjectures from examples that are easily generated by the software. The labs can be assigned as homework or can be used in a structured laboratory setting. Since the software is user-friendly and the laboratories are complete, students and faculty should have no difficulty in using the labs without training.

Most students find that the laboratories provide an enjoyable alternative to the "theorem-proof-example" format of a standard abstract algebra course. At the end of the semester, one student wrote in his evaluation of the course:

I am truly grateful for the laboratory component...Work on the computer helped to make the abstract theory more concrete... One of the best things about the labs was that we formed our own conjectures about the patterns we saw...I believe that the progression of (1) lab, (2) conjecture, (3) class discussion, and (4) proof was highly beneficial in gaining understanding of the abstract material of the course.

Table of Contents: 1. Groups and Geometry; 2. Cayley Tables; 3. Cyclic Groups and Cyclic Subgroups; 4. Subgroups and Subgroup Lattices; 5. The Center and Commutator Subgroups; 6. Quotient Groups; 7. Direct Products; 8. The Unitary Groups; 9. Composition Series; 10. Introduction to Endomorphisms; 11. The Inner Automorphisms of a Group; 12. The Kernel of an Endomorphism; 13. The Class Equation; 14. Conjugate Subgroups; 15. The Sylow Theorems; Appendix A. Table Generation Menu of *Exploring Small Groups (ESG)*; Appendix B. Sample Library of *ESG*; Appendix C. Group Library of *ESG*; Appendix D. Group Properties Menu

Exploring Small Groups, the software packaged with this lab manual, is on a 3½" DD PC compatible disk. This is a DOS program that can be run in Windows. The software was developed by Ladnor Geissinger, University of North Carolina at Chapel Hill.

112 pp., Paperbound, 1996
ISBN 0-88385-705-7
List: $22.00 MAA Member: $16.00
Catalog Code: LABEJR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
PO Box 91112, Washington, DC 20090-1112
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

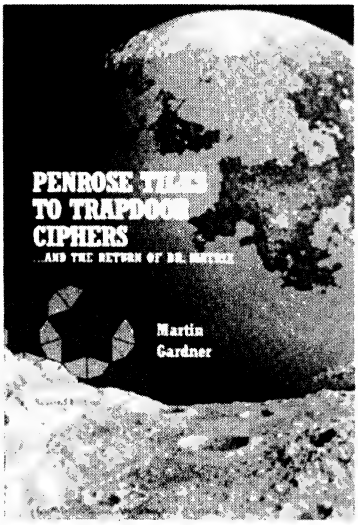
State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	LABE/JR	______	______
______	______________	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________



Penrose Tiles to Trapdoor Ciphers

... and the Return of Dr. Matrix

Martin Gardner
A reissue of another Gardner classic
Series: Spectrum

The MAA is proud to reissue Martin Gardner's *Penrose Tiles to Trapdoor Ciphers*, printed with a new bibliography, corrections to the text, and a postscript from the author. *Penrose Tiles* assembles a collection of Gardner's "Mathematical Games" columns from *Scientific American* that include many of the problems, puzzles and paradoxes that have earned him a reputation as a master mathematical magician.

Included here are chapters on Conway's surreal numbers, Mandelbrot's fractals, and Smullyan's logic puzzles, as well as puzzlers dealing with hyperbolas, negative numbers, pool-ball triangles, and Penrose tiles and trapdoor ciphers. And of course, you can read of the return of Dr. Irvine Joshua Matrix, (famed numerologist and CIA operative), one of Martin Gardner's oldest fictional friends.

Read what reviewers have said about *Penrose Tiles to Trapdoor Ciphers* ...

The scope is extraordinary ... Those fortunate enough to have encountered Gardner's columns in their original appearance can look for personal bonuses of reminiscence as they read this book ... Gardner is one of history's great figures of recreational mathematics. —New Scientist

Penrose Tiles to Trapdoor Ciphers is invaluable to those interested in recreational mathematics and should enlighten those who consider such activity to be difficult or boring. —The Mathematics Teacher.

No popular mathematical writer has ever matched Gardner's breadth and richness of knowledge and clarity of style, and this book is up to his usual unsurpassable standard. —American Scientist

Catalog Code: TILES/JR97
312 pp., Paperbound, 1997, ISBN 0-88385-521-6
List: $27.95 MAA Member: $21.95

Phone in Your Order Now! ☎ ***1-800-331-1622***

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders**: $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ______________________

Address ______________________

City ______________ State _____ Zip ________

Phone ______________________

QTY.	CATALOG CODE	PRICE	AMOUNT
_______	TILES/JR97	_______	_______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling _______

TOTAL _______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

SPRINGER FOR MATHEMATICS

ALEXANDER J. HAHN, University of Notre Dame

BASIC CALCULUS

From Archimedes To Newton To Its Role In Science

This fresh approach to basic calculus is driven in large measure by the responses of the masters of mathematics to the important problems of their day. The text is written in modern style and notation, with an emphasis on the pedagogical factor throughout. It develops calculus from within the relevant historical context, and is suitable for a variety of audiences, from arts and letters to science students. Over 600 exercises let students sharpen their mathematical thinking and skills.

Part 1: From Archimedes to Newton: The Greeks Measure the Universe • Ptolemy and the Dynamics of the Universe • Archimedes Measures Area • A New Astronomy and a New Geometry • The Calculus of Leibnitz • The Calculus of Newton • The Principia.

Part 2: Calculus and the Sciences:Analysis of Functions • Connections with Statics, Dynamics, and Optics • Basic Functions and Their Graphs • The Exponential Function and the Measurement of Age and Growth • The Calculus of Economics • Integral Calculus: Meaning and Methods • Integral Calculus and the Action of Forces

For additional information please see the website:
http://www.nd.edu:80/~hahn/

1997/APP. 900 PP., 540 ILLUS./HARDCOVER/$59.95
ISBN 0-387-94606-3
TEXTBOOKS IN MATHEMATICAL SCIENCES

HUGH GORDON, State University of New York, Albany, NY

DISCRETE PROBABILITY

Discrete Probability is a textbook, at a post-calculus level, for a first course in probability. Since continuous probability is not treated, discrete probability can be covered in greater depth. The result is a book of special interest to students majoring in computer science as well as those majoring in mathematics. Since calculus is used only occasionally, students who have forgotten calculus can nevertheless easily understand the book. The slow, gentle style and clear exposition will appeal to students. Basic concepts such as counting, independence, conditional probability, random variables, approximation of probabilities, generating functions, random walks and Markov chains are presented with good explanation, many worked exercises, and an abundance of problems. Throughout the book, various comments on the history of the study of probability are inserted. Biographical information about some of the famous contributors to probability such as Fermat, Pascal, the Bernoullis, DeMoivre, Bayes, Laplace, Poisson, Markov, and many others, is presented. This volume will appeal to a wide range of readers and be useful in many undergraduate programs.

1997/APP. 256 PP./HARDCOVER/$39.95 ISBN 0-387-98227-2
UNDERGRADUATE TEXTS IN MATHEMATICS

Forthcoming —

ELIAS DEEBA and **ANANDA GUNAWARDENA,** both of University of Houston-Downtown, TX

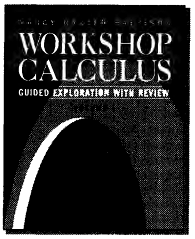
INTERACTIVE LINEAR ALGEBRA WITH MAPLE V

1997/APP. 288 PP./$45.95 (TENT.)/SOFTCOVER/ISBN 0-387-98240-X
TEXTBOOKS IN MATHEMATICAL SCIENCES

NANCY BAXTER HASTINGS, Dickinson College, Carlisle, PA

WORKSHOP CALCULUS

Guided Exploration with Review



Workshop Calculus integrates a review of basic pre-calculus concepts with the study of concepts encountered in a traditional first semester calculus course: functions, limits, derivatives, integrals, and an introduction to integration techniques. This two-course sequence is designed for students who are not prepared to enter Calculus I, but who need to develop mathematical skills for further study in the social sciences, natural sciences, or mathematics. Essential elements of Workshop Calculus include the emphasis on applications to enhance student motivation and the use of computers and graphing calculators to help explore mathematical ideas.

Contents: Volume 1: Functions • Function Construction • Function Classes • Limits • Derivatives and Integrals: First Pass • Volume 2: Derivatives and Integrals: First Pass • Derivatives: A Calculus Approach • Definitive Integrals: A Calculus Approach • Integration Techniques • More Applications

VOL. 1: 1996/424 PP., 42 ILLUS./SOFTCOVER/$29.95
ISBN 0-387-94611-X
VOL. 2: FORTHCOMING 1997
TEXTBOOKS IN MATHEMATICAL SCIENCES

ALAN F. BEARDON, University of Cambridge, UK

LIMITS

A New Approach to Real Analysis

This book is intended as an undergraduate text on real analysis, covering all the standard material such as sequences, infinite series, continuity, differentiation, and integration, together with worked examples and exercises. Despite the fact that there are numerous books on analysis, by unifying and simplifying all the various notions of limit, the author has successfully presented a unique and novel approach to the subject matter, which has not previously appeared in book form. The author defines what is meant by a limit just once, and all of the subsequent limiting processes will be seen as special cases of this one definition. Accordingly, the subject matter attains a unity and coherence that is missing in the traditional approach. Students will be able to fully appreciate and understand the common source of the topics they are studying while also realizing that they are "variations on a theme" rather than essentially different topics, and therefore, will gain a better understanding of the subject.

Contents: Sets, Relations and Directions • Real and Complex Numbers • Limits • Four Basic Results • Infinite Series • Periodic Functions • Sequences • Limits and Continuity • Derivatives • Integration • π, γ, e and n!

1997/APP. 168 PP./HARDCOVER $34.95/ISBN 0-387-98274-4
UNDERGRADUATE TEXTS IN MATHEMATICS

For our latest textbooks please visit us at
{http://www.springer-ny.com/math/text_books/}

Springer
http:www.springer-ny.com

Order Today!

• **CALL:** 1-800-SPRINGER or Fax: (201)-348-4505 • **WRITE:** Springer-Verlag New York, Inc., Dept. S297, PO Box 2485,Secaucus, NJ 07096-2485
• **VISIT:** Your local technical bookstore • **E-MAIL:** orders@springer-ny.com
• **INSTRUCTORS:** Call or write for info on textbook exam copies • *Your 30-Day return privilege is always guarenteed!* 8-9/97 Reference: S297

THE AMERICAN MATHEMATICAL
MONTHLY

Volume 104, Number 8 October 1997

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and Notes should be sent to the Editor:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send your email address and 3 copies of the complete manuscript (including all figures with captions and lettering), typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper.

Letters to the Editor on any topic are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

See the MONTHLY section of MAA Online for current information such as contents of issues, descriptive summaries of forthcoming articles, tips for authors, and preparation of manuscripts in TEX:

http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all problems/solutions material, typewritten on only one side of the paper.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Mr. JOE WATSON, Advertising Manager

Change of address, missing issues inquiries, and other subscription correspondence:
MAA Service Center
maahq@maa.org

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

Areas and Intersections in Convex Domains

Norbert Peyerimhoff

1. INTRODUCTION. The following considerations were inspired by my browsing in the sections about "Bertrand's Paradox" and "Buffon's Needle Problem" of the nice introductory book *The Pleasures of Probability* [**4**]. Both sections discuss problems of an ancient field of mathematics called *geometric probability.*

In some of the subsequent nights I couldn't immediately fall asleep and started to play around with randomly chosen line segments in a bounded convex domain and the probability that they intersect. To avoid ambiguities I have to explain that a straight line segment (A, B) is to be chosen randomly by an independent and uniform random choice of its endpoints A and B in the convex domain. Of course, there are many other methods to choose line segments, which can yield different results for this probability.

Generally, two line segments (A, B) and (C, D) intersect only if they are the diagonals of a convex quadrilateral spanned by the points A, B, C, D. If one of the points A, B, C, D lies properly inside the convex hull of the other three the two line segments cannot intersect. This simple observation connects our problem with an old problem posed by Sylvester more than a century ago: *Given a bounded convex domain $K \subset \mathbb{R}^2$, what is the probability p_K that four points independently chosen in K span a convex quadrilateral*? The connection between these apparently different problems was also important in the recent Monthly article [**12**], which proved that two fundamental constants of planar geometry and geometric probability are equal. The considerations of our article are also related to Sylvester's problem.

In Section 2 we derive a somewhat unexpected relationship between areas of particular subsets of an arbitrary bounded convex domain K by interpreting Sylvester's probability p_K in two different ways. It is a good example of creative interplay between geometry and probability.

Section 3 deals with a higher-dimensional analogue of the original question: *Assume a triangle and a line segment are chosen at random in B^3, the 3-dimensional unit ball. The triangle is determined by choosing three spanning vertices and the line segment is determined by choosing its end points. What is the probability that they intersect*? The solution to this problem is based on two results. We use a theorem of Kingman, which treats a generalization of Sylvester's problem in higher dimensions. We also use a simple case of a fundamental theorem in convex geometry called *Radon's theorem.* It states that any set of $n + 2$ points in $\mathbb{R}^n$ can always be partitioned into two subsets V_1, V_2 such that the convex hulls of V_1 and V_2 intersect.

Henceforth, $\mathrm{conv}(X_1, \dots, X_k)$ denotes the convex hull of the points $X_1, \dots, X_k \in \mathbb{R}^n$ and $\mathrm{aff}(X_1, \dots, X_k)$ denotes the smallest affine subspace containing the points $X_1, \dots, X_k$. We say the points $X_1, \dots, X_k \in \mathbb{R}^n$ are in *general position*, if any affine subspace $P \subset \mathbb{R}^n$ of dimension $r < n$ contains at most $r + 1$ points. The n-dimensional unit ball is denoted by B^n.

2. A STRANGE RELATIONSHIP BETWEEN AREAS AND THE PROBLEM OF SYLVESTER. For the following considerations we fix a convex bounded domain $K \subset \mathbb{R}^2$ and introduce two subsets $\triangle X_1X_2X_3$ and $\Diamond X_1X_2X_3$ associated with three points X_1, X_2, and X_3 of K (see the shaded domains in Figure 1). Let us compare the behavior of area($\triangle X_1X_2X_3$) and area($\Diamond X_1X_2X_3$) as the points X_1, X_2, X_3 move around. The shape of K influences area($\triangle X_1X_2X_3$) only by forcing the points X_1, X_2, X_3 to stay inside K. On the other hand, area($\Diamond X_1X_2X_3$) is very sensitive to the shape of K since $\Diamond X_1X_2X_3$ and K have common pieces of boundary. Moreover, moving the points X_1, X_2, X_3 closer to the boundary of K generally increases area($\triangle X_1X_2X_3$) and decreases area($\Diamond X_1X_2X_3$).

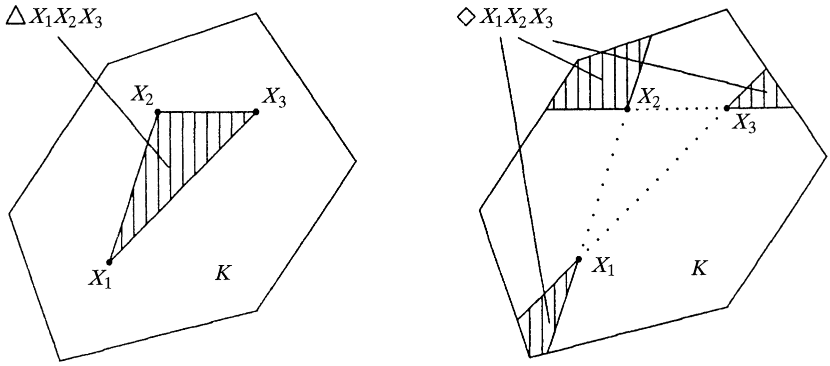


Figure 1

Although both areas behave very differently, there is a surprising connection between them after taking averages. Independent of the shape of K, the average area of $\Diamond X_1X_2X_3$ is exactly three times the average area of $\triangle X_1X_2X_3$, as described precisely in the following proposition.

Proposition 1. *Let $K \subset \mathbb{R}^2$ be a given bounded convex domain. Then*

$$\frac{1}{\operatorname{area}(K)^3}\int_{K\times K\times K} \operatorname{area}(\Diamond X_1X_2X_3)\, dX_1 dX_2 dX_3$$
$$= \frac{3}{\operatorname{area}(K)^3}\int_{K\times K\times K} \operatorname{area}(\triangle X_1X_2X_3)\, dX_1 dX_2 dX_3. \quad (1)$$

The two normalized integrals in (1) *are, respectively, what we mean by the average areas of $\Diamond X_1X_2X_3$ and $\triangle X_1X_2X_3$.*

The main idea of the proof is to express a particular probability in two different ways. Let

$$p_{\triangle} := \begin{array}{l}\text{probability that the convex hull of four independently}\\ \text{chosen points } X_1, \dots, X_4 \in K \text{ is a triangle.}\end{array} \quad (2)$$

Bertrand's paradox tells us that one has to be careful in using probabilities for geometric settings. We avoid these difficulties by restricting ourselves only to independent and uniform choices of points in the convex bounded domain K. Any choice of 4 points corresponds to an element in the product space $K \times K \times K \times K$. By using Fubini's theorem we can reduce the situation to the probability that a randomly chosen point lies inside some measurable subset $A \subset K$. This probability

is given by area$(A)/$area(K), which corresponds to the quotient

$$\frac{\#\text{ good cases}}{\#\text{ all cases}}$$

in the case of finite sets. This relationship between probabilities and areas is crucial in the proof.

Proof: Let $X_1, \dots, X_4$ be four independently chosen points in K. We can restrict our considerations to points in general position since the event "X_1, X_2, X_3, X_4 are not in general position" has probability zero. If conv(X_1, X_2, X_3, X_4) is a triangle, one of the four points must lie inside the convex hull of the other three. This interior point is uniquely determined. For $j = 1, \dots, 4$, let p_j denote the probability of the event "conv(X_1, X_2, X_3, X_4) is a triangle *and* X_j is the interior point." By symmetry, note that $p_1 = p_2 = p_3 = p_4$. Then the probability $p_\diamond$ defined in (2) satisfies $p_\triangle = p_1 + p_2 + p_3 + p_4 = 4p_4$.

For fixed $X_1, X_2, X_3 \in K$ the probability of the event "$X_4 \in$ conv(X_1, X_2, X_3)" (X_4 chosen randomly in K) is given by area$(\triangle X_1X_2X_3)/$area(K). Since $X_1, X_2, X_3 \in K$ are also chosen randomly we have to integrate over all possible triples $(X_1, X_2, X_3) \in K \times K \times K$

$$p_4 := \frac{1}{\text{area}(K)^3} \int_{K\times K\times K} \frac{\text{area}(\triangle X_1X_2X_3)}{\text{area}(K)}\, dX_1 dX_2 dX_3,$$

and consequently,

$$p_\triangle = \frac{4}{\text{area}(K)^3} \int_{K\times K\times K} \frac{\text{area}(\triangle X_1X_2X_3)}{\text{area}(K)}\, dX_1 dX_2 dX_3. \tag{3}$$

In a second approach to $p_\triangle$, we again assume $X_1, X_2, X_3 \in K$ to be fixed and ask for the probability of "conv(X_1, X_2, X_3, X_4) is a triangle" if X_4 is chosen randomly in K. This happens exactly if X_4 lies inside the shaded domain of Figure 2, which is $\triangle X_1X_2X_3 \cup \diamond X_1X_2X_3$. Thus this probability is given by

$$\frac{\text{area}(\triangle X_1X_2X_3) + \text{area}(\diamond X_1X_2X_3)}{\text{area}(K)}.$$

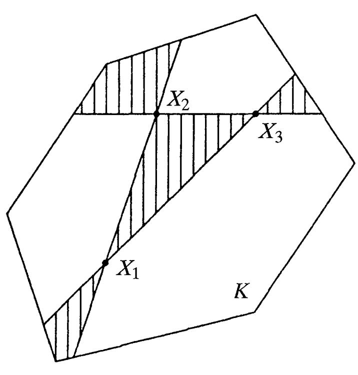


Figure 2

We again obtain $p_\triangle$ by integrating this probability over all possible choices $(X_1, X_2, X_3) \in K \times K \times K$:

$$p_\triangle = \frac{1}{\operatorname{area}(K)^3} \int_{K\times K\times K} \frac{\operatorname{area}(\triangle X_1X_2X_3) + \operatorname{area}(\Diamond X_1X_2X_3)}{\operatorname{area}(K)}\, dX_1 dX_2 dX_3. \tag{4}$$

Comparison of (3) and (4) yields the statement of the proposition. ■

Dropping the condition area$(K) < \infty$ (which is guaranteed by the boundedness of K) leads to serious difficulties. For example, the probability that the convex hull of four randomly and independently chosen points in the whole plane is a triangle is not well-defined.

Obviously, $p_\triangle$ and Sylvester's probability p_K mentioned in the introduction are probabilities of complementary events, hence $p_K = 1 - p_\triangle$. Some further remarks about Sylvester's probability are in order. Invariance of p_K under affine transformations of K implies that its value is the same for all triangles as well as for all ellipses. For these cases p_K is known [**7**, pp. 44–46]:

$$p_K = \begin{cases} \dfrac{2}{3} & \text{if } K \text{ is a triangle} \\ 1 - \dfrac{35}{12\pi^2} & \text{if } K \text{ is an ellipse.} \end{cases}$$

W. Blaschke [**1**, §24, §25] gave the first rigorous proof that the triangle and the ellipse are the extremal cases for p_K and that $2/3 \le p_K \le 1 - 35/12\pi^2$ for any bounded convex domain K. Blaschke's proof is surely very appealing to readers interested in geometric ideas.

I would like to mention two generalizations of Sylvester's problem. J. F. C. Kingman investigated the higher-dimensional analogue of Sylvester's problem and solved it for the case $K = B^n$ (see [**5**, Theorem 7]). The extremality property of ellipses for p_K was generalized to higher dimensions by H. Groemer [**3**]:

Theorem 2. (Kingman, Groemer) *The probability $p_{\triangle,n}$ that the convex hull of $n + 2$ randomly chosen points in the unit ball $B^n \subset \mathbb{R}^n$ has $n + 1$ vertices is given by*

$$p_{\triangle,n} = (n+2)\binom{n+1}{\frac{n+1}{2}}^{n+1} \Bigg/ \left(\binom{(n+1)^2}{\frac{(n+1)^2}{2}} 2^n\right). \tag{5}$$

The corresponding probability for any bounded convex domain $K \subset \mathbb{R}^n$ is never greater than the value given in (5).

The binomial coefficients in (5) are defined by

$$\binom{a}{b} = \frac{\Gamma(a+1)}{\Gamma(b+1)\Gamma(a-b+1)}$$

if not both parameters a, b are integers. Since $\Gamma(1/2) = \sqrt{\pi}$ we obtain $p_{\triangle,2} = 35/12\pi^2$. For $n = 3$ we obtain $p_{\triangle,3} = 9/143$, which will be used in the next section.

Another generalization of Sylvester's problem was investigated much more recently by P. Valtr: *What is the probability $p_K(n)$ that n randomly chosen points in a*

fixed bounded convex domain $K \subset \mathbb{R}^2$ are the vertices of a convex n-gon? Sylvester's problem is then the special case $n = 4$. Valtr obtained explicit formulas for this more general problem if K is a parallelogram or a triangle [**13, 14**]:

$$p_K(n) = \begin{cases} \dfrac{2^n(3n-3)!}{((n-1)!)^3(2n)!} & \text{if } K \text{ is a triangle} \\[2ex] \left(\dfrac{\binom{2n-2}{n-1}}{n!}\right)^2 & \text{if } K \text{ is a parallelogram.} \end{cases}$$

Interestingly, Valtr's method is purely combinatorial and does not require any integration. I recommend [**6**], [**11**, pp. 63–65], and [**15**] for further information about this subject.

3. ABOUT THE PROBABILITY OF AN INTERSECTION. This section gives an answer to the following question: *Let us assume that five points X_1, X_2, X_3, X_4, X_5 are chosen independently and uniformly at random in the unit ball B^3. The convex hull of X_1, X_2, X_3 is generally a triangle $\triangle X_1X_2X_3$ and the convex hull of X_4, X_5 is generally a straight line segment (X_4, X_5). What is the probability of the event "$\triangle X_1X_2X_3 \cap (X_4, X_5) \neq \emptyset$"?*

The following simple argument shows that this probability must be less than $1/2$. Let $X_1, X_2, X_3 \in B^3$ be randomly chosen points in general position. Since $\triangle X_1X_2X_3 \subset \mathrm{aff}(X_1, X_2, X_3)$ the probability of "$\triangle X_1X_2X_3 \cap (X_4, X_5) \neq \emptyset$" is at most the probability of the event "$\mathrm{aff}(X_1, X_2, X_3) \cap (X_4, X_5) \neq \emptyset$." The plane $\mathrm{aff}(X_1, X_2, X_3)$ cuts the ball B^3 into two disjoint subsets S_1, S_2. Let p and q denote the probabilities that a randomly chosen point $Y \in B^3$ lies in S_1 or S_2, respectively. Assuming the plane $\mathrm{aff}(X_1, X_2, X_3)$ to be chosen and fixed, $2pq$ is the probability that the line segment (X_4, X_5) of two randomly chosen points $X_4, X_5 \in B^3$ intersects this plane (this is the event "X_4 and X_5 fall into different subsets S_1 and S_2"). From $p + q = 1$ we conclude that $2pq \leq 1/2$. By averaging over all choices of X_1, X_2, X_3 we obtain a final probability $\leq 1/2$, but we shall see that this upper bound is very crude.

To determine the exact probability, we first investigate the possible shapes of the convex hull of five randomly chosen points $X_1, X_2, X_3, X_4, X_5 \in \mathbb{R}^3$ in general position. It turns out that $\mathrm{conv}(X_1, \dots, X_5)$ is either a double pyramid over a triangle with tips in different directions, or it is a tetrahedron. This statement seems obvious but the author doesn't know a three-line proof. However, it is an easy consequence of the following basic theorem in convex geometry:

Theorem 3. (Radon). *If $V = \{X_1, X_2, \dots, X_{n+2}\}$ is a set of $n + 2$ points of $\mathbb{R}^n$ in general position then there exists a partition of these points into two sets V_1, V_2 with $V_1 \cup V_2 = \{X_1, \dots, X_{n+2}\}$ and $V_1 \cap V_2 = \emptyset$ such that the convex hulls* $\mathrm{conv}(V_1)$ *and* $\mathrm{conv}(V_2)$ *intersect in exactly one point. This "Radon partition" V_1, V_2 is unique in the sense that, for any other partition, the intersection of the corresponding convex hulls is empty. Here we identify partitions that are obtained by interchanging V_1 and V_2.*

Radon's theorem can be proved very elegantly with methods of linear algebra (see, e.g., [**8**, pp. 22–24] or exercise 6.0 in [**16**]). The algebraic methods, however, don't illustrate the geometric intuition of the theorem. Readers interested in a geometric proof should look into [**10**] and [**9**]. A thorough treatment of the historical background of the theorem and related results can be found in [**2**].

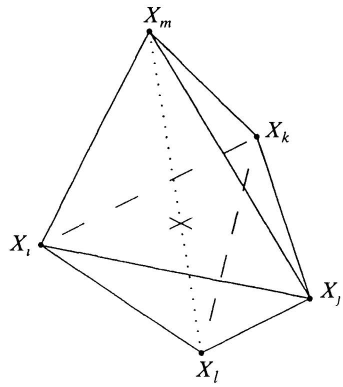


Figure 3

Corollary 4. *Assume that there are five points* $X_1, X_2, \ldots, X_5 \in \mathbb{R}^3$ *in general position whose convex hull has five vertices. Then one can always choose a partition* $V_1 = \{ X_i, X_j, X_k\}$ *and* $V_2 = \{X_l, X_m\}$ *such that*

$$\Delta X_i X_j X_k \cap (X_l, X_m) \neq \emptyset.$$

This partition is unique in the sense of Theorem 3.

Proof: Since the convex hull P has five vertices, none of the points $X_1, \ldots, X_5$ lies inside the convex hull of the other four. Hence the unique partition V_1, V_2 in Theorem 3 must consist of 3 and 2 points. ■

Figure 3 illustrates a case in which the convex hull of $X_1, \ldots, X_5$ has five vertices. To confirm the uniqueness of the partition, observe that the triangle and the segment defined by the partition $V_1 = \{X_l, X_i, X_m\}$, $V_2 = \{X_j, X_k\}$ don't intersect.

If $X_1, \ldots, X_5$ are in general position and $P := \mathrm{conv}(X_1, \ldots, X_5\}$ is a tetrahedron, the unique Radon partition is

$$V_1 := \{\text{the four vertices of } P\},$$

$$V_2 := \{\text{remaining fifth point in the interior of } P\},$$

and any other partition yields an empty intersection. Consequently, the intersection of a triangle spanned by any three of the points $X_1, \ldots, X_5$ with the line segment spanned by the remaining two points is always empty.

With this information in hand we are able to calculate the probability of the event "$\triangle X_1 X_2 X_3 \cap (X_4, X_5) \neq \emptyset$" for randomly chosen points $X_1, \ldots, X_5 \in B^3$ as

$$p_{32} := \frac{1}{\mathrm{vol}(B^3)^5} \int_{B^3 \times \cdots \times B^3} m_{32}(X_1, \ldots, X_5)\, dX_1 \cdots dX_5 \tag{7}$$

where the function in the integrand is

$$m_{32}(X_1, X_2, X_3, X_4, X_5) := \begin{cases} 1 & \text{if } \triangle X_1 X_2 X_3 \cap (X_4, X_5) \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

Let us denote by $\mathscr{S}_5$ the group of permutations of the numbers $1, 2, 3, 4, 5$. For any five points $X_1, \dots, X_5$ in general position, we shall show later that Corollary 4 ensures that

$$\sum_{\sigma \in \mathscr{S}_5} m_{32}(X_{\sigma_1}, X_{\sigma_2}, X_{\sigma_3}, X_{\sigma_4}, X_{\sigma_5}) = \begin{cases} 12 & \text{if conv}(X_1, \dots, X_5) \text{ has 5 vertices} \\ 0 & \text{otherwise.} \end{cases}$$

In the right-hand side of (7) we can permute the variables $X_1, X_2, \dots, X_5$ without changing its value. If we rewrite (7) in the following more complicated way

$$p_{32} = \frac{1}{\text{vol}(B^3)^5} \frac{1}{5!} \sum_{\in \mathscr{S}_5} \int_{B^3 \times \cdots \times B^3} m_{32}(X_{\sigma_1}, \dots, X_{\sigma_5})\, dX_1 \cdots dX_5,$$

we are able to apply (8) and obtain

$$p_{32} = \frac{12}{5!} \frac{1}{\text{vol}(B^3)^5} \int_{B^3 \times \cdots \times B^3} k_5(X_1, \dots, X_5)\, dX_1 \cdots dX_5, \tag{9}$$

where k_5 is 1, if conv$(X_1, \dots, X_5)$ has 5 vertices, and 0, otherwise. The normalized integral in (9) is the probability that the convex hull of 5 randomly chosen points in B^3 has 5 vertices, which is $1 - p_{\triangle,3}$, and Theorem 2 ensures that $p_{\triangle,3} = 9/143$. Consequently, the answer to the question posed at the beginning of this section is:

Proposition 5. *Assume that $X_1, \dots, X_5$ are chosen randomly in the unit ball $B^3 \subset \mathbb{R}^3$. The probability of the event "$\triangle X_1 X_2 X_3 \cap (X_4, X_5) \neq \emptyset$" is*

$$p_{32} = \frac{1}{10} \frac{134}{143} \approx 0.0937\dots$$

The corresponding probability for any bounded convex set $K \subset \mathbb{R}^3$ is no less than p_{32}.

It remains to prove (8): If $X_1, \dots, X_5$ are in general position, conv$(X_1, \dots, X_5)$ has either 5 or 4 vertices. In the latter case, $m_{32}(X_{\sigma_1}, \dots, X_{\sigma_5}) = 0$ for any permutation $\sigma \in \mathscr{S}_5$. If conv$(X_{\sigma_1}, \dots, X_{\sigma_5})$ has 5 vertices we conclude from Corollary 4 that

$$m_{32}(X_{\sigma_1}, \dots, X_{\sigma_5}) = 1$$

if and only if $\{\sigma_1, \sigma_2, \sigma_3\} = \{i, j, k\}$ and $\{\sigma_4, \sigma_5\} = \{l, m\}$. This is true for exactly $3!2! = 12$ of all the permutations. ∎

4. FINAL REMARKS. Proposition 1 can be generalized to higher dimensions. For example, in 3 dimensions any four points $X_1, \dots, X_4 \in K$ in general position span a simplex S. The planes containing the two dimensional faces of S cut the domain K into 15 subsets. In this case there exists a connection between the area of the simplex spanned by $X_1, \dots, X_4$ and the total area of the four subsets that touch the simplex in exactly one vertex.

I would like to mention two problems related to the considerations of Section 3. The first problem is an easy exercise whereas the second might be a difficult research problem. Let p_{kl} denote the probability that conv$(X_1, \dots, X_k)$ and conv$(Y_1, \dots, Y_l)$ intersect where $X_1, \dots, X_k$ and $Y_1, \dots, Y_l$ are randomly chosen points in the unit ball B^{k+l-2}.

Problem 1. Determine the probability p_{22} that two randomly chosen segments in B^2 intersect. Check your result by running a *Monte Carlo test.*

Problem 2. It is possible to calculate the probabilities $p_{1k} = p_{k1}$ for all $k \geq 2$, p_{22}, and $p_{23} = p_{32}$ by combining Kingman's and Radon's theorem. This simple method fails to work for all the other probabilities p_{kl}. Find new methods to calculate some of the other probabilities.

Last, but not least, I want to express my gratitude to Edson de Faria, Reinhard Hermann, Richard Isaac, Lucian Man, Mechthild Stoer, Pavel Valtr, and Günther M. Ziegler for many constructive comments. Luiz Magalhães made helpful suggestions for improving the figures in this article. I am also thankful to the Deutsche Forschungsgemeinschaft (DFG) for its financial support.

REFERENCES

1. W. Blaschke, *Vorlesungen über Differentialgeometrie, II Affine Differentialgeometrie*, Grundlehren der Mathematischen Wissenschaften, Springer, Berlin, 1923.
2. J. Eckhoff, *Helly, Radon, and Carathéodory type theorems*, Chapter 2.1 in: P.M. Gruber and J.M. Wills (eds.), *Handbook of Convex Geometry*, North-Holland 1993, Amsterdam, pp. 389–448.
3. H. Groemer, On some mean values associated with a randomly selected simplex in a convex set, *Pacific J. Math.* **45** (1973), 525–533.
4. R. Isaac, *The Pleasures of Probability*, Springer, Readings in Mathematics, New York, 1995.
5. J. F. C. Kingman, Random secants of a convex body, *J. Appl. Prob.* **6** (1969), 660–672.
6. V. Klee, What is the expected volume of a simplex whose vertices are chosen at random from a given convex body?, this *Monthly* **76** (1969), 286–288.
7. M. G. Kendall and P. A. P. Moran, *Geometrical Probability*, Griffin's statistical monographs & courses, Hafner, New York, 1963.
8. P. McMullen and G. C. Shepard, *Convex Polytopes and the Upper Bound Conjecture*, London Math. Soc. Lecture Note Series 3, Cambridge University Press, London, 1971.
9. B. B. Peterson, The geometry of Radon's theorem, this *Monthly* **79** (1972), 949–963.
10. R. Rado, Theorems on the intersection of convex sets of points, *J. London Math. Soc.* **27** (1952), 320–328.
11. L. A. Santaló, *Integral Geometry and Geometric Probability*, Addison-Wesley, Reading, Massachusetts, 1976.
12. E. R. Scheinerman and H. S. Wilf, The Rectilinear Crossing Number of a complete graph and Sylvester's "Four Point Problem" of Geometric Probability, this *Monthly* **101** (1994), 939–943.
13. P. Valtr, Probability that n random points are in convex position, *Discrete Comput. Geometry* **13** (1995), 637–643.
14. P. Valtr, The probability that n random points in a triangle are in convex position, to appear in *Combinatorica*.
15. W. Weil and J. Wieacker, *Stochastic Geometry*, Chapter 5.2 in: P.M. Gruber and J.M. Wills (eds.), *Handbook of Convex Geometry*, North-Holland, Amsterdam, 1993, pp. 1391–1438.
16. G. M. Ziegler, *Lectures on Polytopes*, Springer, Graduate Texts in Mathematics 152, New York, 1994.

Mathematisches Institut
Universität Basel
Rheinsprung 21
CH-4051 Basel
peyerim@math.unibas.ch

Newman's Short Proof of the Prime Number Theorem

D. Zagier

Dedicated to the Prime Number Theorem on the occasion of its 100th birthday

The prime number theorem, that the number of primes $\leq x$ is asymptotic to $x/\log x$, was proved (independently) by Hadamard and de la Vallée Poussin in 1896. Their proof had two elements: showing that Riemann's zeta function $\zeta(s)$ has no zeros with $\Re(s) = 1$, and deducing the prime number theorem from this. An ingenious short proof of the first assertion was found soon afterwards by the same authors and by Mertens and is reproduced here, but the deduction of the prime number theorem continued to involve difficult analysis. A proof that was elementary in a technical sense—it avoided the use of complex analysis—was found in 1949 by Selberg and Erdös, but this proof is very intricate and much less clearly motivated than the analytic one. A few years ago, however, D. J. Newman found a very simple version of the Tauberian argument needed for an analytic proof of the prime number theorem. We describe the resulting proof, which has a beautifully simple structure and uses hardly anything beyond Cauchy's theorem.

Recall that the notation $f(x) \sim g(x)$ ("f and g are asymptotically equal") means that $\lim_{x \to \infty} f(x)/g(x) = 1$, and that $O(f)$ denotes a quantity bounded in absolute value by a fixed multiple of f. We denote by $\pi(x)$ the number of primes $\leq x$.

Prime Number Theorem. $\pi(x) \sim \dfrac{x}{\log x}$ as $x \to \infty$.

We present the argument in a series of steps. Specifically, we prove a sequence of properties of the three functions

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}, \quad \Phi(s) = \sum_p \frac{\log p}{p^s}, \quad \vartheta(x) = \sum_{p \leq x} \log p \quad (s \in \mathbb{C},\ x \in \mathbb{R});$$

we always use p to denote a prime. The series defining $\zeta(s)$ (the Riemann zeta-function) and $\Phi(s)$ are easily seen to be absolutely and locally uniformly convergent for $\Re(s) > 1$, so they define holomorphic functions in that domain.

(I). $\zeta(s) = \prod_p (1 - p^{-s})^{-1}$ *for* $\Re(s) > 1$.

Proof: From unique factorization and the absolute convergence of $\zeta(s)$ we have

$$\zeta(s) = \sum_{r_2, r_3, \cdots \geq 0} (2^{r_2} 3^{r_3} \cdots)^{-s} = \prod_p \Big(\sum_{r \geq 0} p^{-rs} \Big) = \prod_p \frac{1}{1 - p^{-s}} \quad (\Re(s) > 1).$$

(II). $\zeta(s) - \dfrac{1}{s-1}$ *extends holomorphically to* $\Re(s) > 0$.

Proof: For $\Re(s) > 1$ we have

$$\zeta(s) - \frac{1}{s-1} = \sum_{n=1}^{\infty} \frac{1}{n^s} - \int_1^{\infty} \frac{1}{x^s}\,dx = \sum_{n=1}^{\infty} \int_n^{n+1} \left(\frac{1}{n^s} - \frac{1}{x^s} \right) dx.$$

The series on the right converges absolutely for $\Re(s) > 0$ because

$$\left| \int_n^{n+1} \left(\frac{1}{n^s} - \frac{1}{x^s} \right) dx \right| = \left| s \int_n^{n+1} \int_n^x \frac{du}{u^{s+1}}\,dx \right| \le \max_{n \le u \le n+1} \left| \frac{s}{u^{s+1}} \right| = \frac{|s|}{n^{\Re(s)+1}}$$

by the mean value theorem.

(III). $\vartheta(x) = O(x)$.

Proof: For $n \in \mathbb{N}$ we have

$$2^{2n} = (1+1)^{2n} = \binom{2n}{0} + \cdots + \binom{2n}{2n} \ge \binom{2n}{n} \ge \prod_{n < p \le 2n} p = e^{\vartheta(2n) - \vartheta(n)}$$

and hence, since $\vartheta(x)$ changes by $O(\log x)$ if x changes by $O(1)$, $\vartheta(x) - \vartheta(x/2) \le Cx$ for any $C > \log 2$ and all $x \ge x_0 = x_0(C)$. Summing this over $x, x/2, \ldots, x/2^r$, where $x/2^r \ge x_0 > x/2^{r+1}$, we obtain $\vartheta(x) \le 2Cx + O(1)$.

(IV). $\zeta(s) \ne 0$ *and* $\Phi(s) - 1/(s-1)$ *is holomorphic for* $\Re(s) \ge 1$.

Proof: For $\Re(s) > 1$, the convergent product in (I) implies that $\zeta(s) \ne 0$ and that

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_p \frac{\log p}{p^s - 1} = \Phi(s) + \sum_p \frac{\log p}{p^s(p^s - 1)}.$$

The final sum converges for $\Re(s) > \frac{1}{2}$, so this and (II) imply that $\Phi(s)$ extends meromorphically to $\Re(s) > \frac{1}{2}$, with poles only at $s = 1$ and at the zeros of $\zeta(s)$, and that, if $\zeta(s)$ has a zero of order μ at $s = 1 + i\alpha$ ($\alpha \in \mathbb{R}$, $\alpha \ne 0$) and a zero of order ν at $1 + 2i\alpha$(so $\mu, \nu \ge 0$ by (II)), then

$$\lim_{\epsilon \searrow 0} \epsilon \Phi(1+\epsilon) = 1,\ \lim_{\epsilon \searrow 0} \epsilon \Phi(1 + \epsilon \pm i\alpha) = -\mu, \text{ and } \lim_{\epsilon \searrow 0} \epsilon \Phi(1 + \epsilon \pm 2i\alpha) = -\nu.$$

The inequality

$$\sum_{r=-2}^{2} \binom{4}{2+r} \Phi(1 + \epsilon + ir\alpha) = \sum_p \frac{\log p}{p^{1+\epsilon}} \left(p^{i\alpha/2} + p^{-i\alpha/2} \right)^4 \ge 0$$

then implies that $6 - 8\mu - 2\nu \ge 0$, so $\mu = 0$, i.e., $\zeta(1 + i\alpha) \ne 0$.

(V). $\int_1^{\infty} \frac{\vartheta(x) - x}{x^2}\,dx$ *is a convergent integral.*

Proof: For $\Re(s) > 1$ we have

$$\Phi(s) = \sum_p \frac{\log p}{p^s} = \int_1^{\infty} \frac{d\vartheta(x)}{x^s} = s \int_1^{\infty} \frac{\vartheta(x)}{x^{s+1}}\,dx = s \int_0^{\infty} e^{-st} \vartheta(e^t)\,dt.$$

Therefore (V) is obtained by applying the following theorem to the two functions $f(t) = \vartheta(e^t)e^{-t} - 1$ and $g(z) = \Phi(z+1)/(z+1) - 1/z$, which satisfy its hypotheses by (III) and (IV).

Analytic Theorem. *Let $f(t)$ $(t \geq 0)$ be a bounded and locally integrable function and suppose that the function $g(z) = \int_0^\infty f(t)e^{-zt}\,dt$ $(\Re(z) > 0)$ extends holomorphically to $\Re(z) \geq 0$. Then $\int_0^\infty f(t)\,dt$ exists (and equals $g(0)$).*

(VI). $\vartheta(x) \sim x$.

Proof: Assume that for some $\lambda > 1$ there are arbitrarily large x with $\vartheta(x) \geq \lambda x$. Since ϑ is non-decreasing, we have

$$\int_x^{\lambda x} \frac{\vartheta(t) - t}{t^2}\,dt \geq \int_x^{\lambda x} \frac{\lambda x - t}{t^2}\,dt = \int_1^{\lambda} \frac{\lambda - t}{t^2}\,dt > 0$$

for such x, contradicting (V). Similarly, the inequality $\vartheta(x) \leq \lambda x$ with $\lambda < 1$ would imply

$$\int_{\lambda x}^{x} \frac{\vartheta(t) - t}{t^2}\,dt \leq \int_{\lambda x}^{x} \frac{\lambda x - t}{t^2}\,dt = \int_{\lambda}^{1} \frac{\lambda - t}{t^2}\,dt < 0,$$

again a contradiction for λ fixed and x big enough.

The prime number theorem follows easily from (VI), since for any $\epsilon > 0$

$$\vartheta(x) = \sum_{p \leq x} \log p \leq \sum_{p \leq x} \log x = \pi(x)\log x,$$

$$\vartheta(x) \geq \sum_{x^{1-\epsilon} \leq p \leq xx} \log p \geq \sum_{x^{1-\epsilon} \leq p \leq x} (1-\epsilon)\log x$$

$$= (1-\epsilon)\log x\left[\pi(x) + O(x^{1-\epsilon})\right].$$

Proof of the Analytic Theorem. For $T > 0$ set $g_T(z) = \int_0^T f(t)e^{-zt}\,dt$. This is clearly holomorphic for all z. We must show that $\lim_{T\to\infty} g_T(0) = g(0)$.

Let R be large and let C be the boundary of the region $\{z \in \mathbb{C} \mid |z| \leq R, \Re(z) \geq -\delta\}$, where $\delta > 0$ is small enough (depending on R) so that $g(z)$ is holomorphic in and on C. Then

$$g(0) - g_T(0) = \frac{1}{2\pi i}\int_C (g(z) - g_T(z))e^{zT}\left(1 + \frac{z^2}{R^2}\right)\frac{dz}{z}$$

by Cauchy's theorem. On the semicircle $C_+ = C \cap \{\Re(z) > 0\}$ the integrand is bounded by $2B/R^2$, where $B = \max_{t\geq 0}|f(t)|$, because

$$|g(z) - g_T(z)| = \left|\int_T^\infty f(t)e^{-zt}\,dt\right| \leq B\int_T^\infty |e^{-zt}|\,dt = \frac{Be^{-\Re(z)T}}{\Re(z)} \quad (\Re(z) > 0)$$

and

$$\left|e^{zT}\left(1 + \frac{z^2}{R^2}\right)\frac{1}{z}\right| = e^{\Re(z)T}\cdot\frac{2\Re(z)}{R^2}.$$

Hence the contribution to $g(0) - g_T(0)$ from the integral over C_+ is bounded in absolute value by B/R. For the integral over $C_- = C \cap \{\Re(z) < 0\}$ we look at $g(z)$ and $g_T(z)$ separately. Since g_T is entire, the path of integration for the integral involving g_T can be replaced by the semicircle $C'_- = \{z \in \mathbb{C} \mid |z| = R, \Re(z) < 0\}$, and the integral over C'_- is then bounded in absolute value by $2\pi B/R$

by exactly the same estimate as before since

$$|g_T(z)| = \left|\int_0^T f(t)e^{-zt}\,dt\right| \le B\int_{-\infty}^{T} |e^{-zt}|\,dt = \frac{Be^{-\Re(z)T}}{|\Re(z)|} \quad (\Re(z) < 0).$$

Finally, the remaining integral over C_- tends to 0 as $T \to \infty$ because the integrand is the product of the function $g(z)(1 + z^2/R^2)/z$, which is independent of T, and the function e^{zT}, which goes to 0 rapidly and uniformly on compact sets as $T \to \infty$ in the half-plane $\Re(z) < 0$. Hence $\limsup_{T\to\infty}|g(0) - g_T(0)| \le 2B/R$. Since R is arbitrary this proves the theorem.

Historical remarks. The "Riemann" zeta function $\zeta(s)$ was first introduced and studied by Euler, and the product representation given in (I) is his. The connection with the prime number theorem was found by Riemann, who made a deep study of the analytic properties of $\zeta(s)$. However, for our purposes the nearly trivial analytic continuation property (II) is sufficient. The extremely ingenious proof in (III) is in essence due to Chebyshev, who used more refined versions of such arguments to prove that the ratio of $\vartheta(x)$ to x (and hence also of $\pi(x)$ to $x/\log x$) lies between 0.92 and 1.11 for x sufficiently large. This remained the best result until the prime number theorem was proved in 1896 by de la Vallée Poussin and Hadamard. Their proofs were long and intricate. (A simplified modern presentation is given on pages 41–47 of Titchmarsh's book on the Riemann zeta function [T].) The very simple proof reproduced in (IV) of the non-vanishing of $\zeta(s)$ on the line $\Re(s) = 1$ was given in essence by Hadamard (the proof of this fact in de la Vallée Poussin's first paper had been about 25 pages long) and then refined by de la Vallée Poussin and by Mertens, the version given by the former being particularly elegant. The Analytic Theorem and its use to prove the prime number theorem as explained in steps (V) and (VI) above are due to D. J. Newman. Apart from a few minor simplifications, the exposition here follows that in Newman's original paper [N] and in the expository paper [K] by J. Korevaar.

We refer the reader to P. Bateman and H. Diamond's survey article [B] for a beautiful historical perspective on the prime number theorem.

REFERENCES

[B] P. Bateman and H. Diamond, A hundred years of prime numbers, *Amer. Math. Monthly* **103** (1996), 729–741.

[K] J. Korevaar, On Newman's quick way to the prime number theorem, *Math. Intelligencer* **4**, 3 (1982), 108–115.

[N] D. J. Newman, Simple analytic proof of the prime number theorem, *Amer. Math. Monthly* **87** (1980), 693–696.

[T] E. C. Titchmarsh, *The Theory of the Riemann Zeta Function*, Oxford, 1951.

Max-Planck-Institut für Mathematik
Gottfried-Claren-Straße 26
D-53225 Bonn, Germany
zagier@mpim-bonn.mpg.de

An Exploratory Approach to Kaplansky's Lemma Leads to a Generalized Resultant

David Callan

Let $T: V \to V$ be a linear transformation on a vector space V. If the powers $I, T, T^2, \dots, T^k$ of T are linearly dependent, then surely for every $v \in V$, the vectors $v, Tv, T^2v, \dots, T^kv$ are also linearly dependent. Kaplansky noted that the converse assertion is also true.

Theorem 1. (Kaplansky's Lemma in finite dimensions). *Suppose $T: V \to V$ is a linear transformation on a finite-dimensional vector space V over an arbitrary field F and suppose the set of powers $\{T^i\}_{i=0}^k$ is linearly independent. Then there exists $v \in V$ such that the set $\{T^iv\}_{i=0}^k$ is linearly independent.*

This result actually holds in arbitrary dimensions and the extension from finite to infinite dimensions is fairly straightforward [**6**, p. 64]. Aupetit [**1**] gives an elegant proof when $F = \mathbb{C}$ (valid for any dimensions) using Liouville's theorem from complex analysis. His proof is reproduced, slightly polished, in Prasolov's recent intriguing survey of linear algebra emphasizing modern proofs of classical results [**5**, p. 46].

Henceforth, we assume a finite-dimensional space, so T has a minimal polynomial of degree n, say. The maximum k can be is $n-1$ and so we may as well assume $k = n-1$. Prasolov then gives a slick proof, but only for infinite fields, using minimal vector-annihilating polynomials and their subspaces [**5**, 13.1.2, p. 73]. It relies on the result: a vector space cannot be expressed as a finite set-theoretic union of proper subspaces, and this of course requires the underlying field to be infinite. For arbitrary fields, one can readily construct a suitable vector v using the theory of the rational canonical form [**4**, p. 198]. This says that T acts as a block diagonal matrix, the diagonal blocks being the companion matrices of the invariant factors $d_1(x) \mid d_2(x) \mid \dots \mid d_s(x)$ of T. The vector v consisting of 0's except for a 1 in the first position corresponding to the $d_s(x)$ block has minimal annihilating polynomial $d_s(x)$. This vector v fits the bill since $d_s(x)$ is also the minimal polynomial of T.

But suppose out of curiosity we look for a suitable v using the version of the rational canonical form involving elementary divisors [**3**, p. 262]. Now T acts as a block diagonal matrix whose diagonal blocks are the companion matrices of the elementary divisors $\{p_i(x)^{e_{ij}}\}$ of T. Here the polynomials $p_i(x)$ are irreducible and distinct. Let $f_i(x) = p_i(x)^{r_i}$, $1 \le i \le m$, be the elementary divisors of highest degree, so that $\prod_{i=1}^m f_i(x)$ is the minimal polynomial of T with degree $n = \sum_{i=1}^m \deg f_i$. In fact, it is sufficient to assume these are the *only* elementary divisors; after dealing with this case, v can be augmented with zeros to cover the general case.

Thus we assume T is the n-square matrix $\operatorname{diag}(A_1, A_2, \dots, A_m)$ where A_i is the companion matrix of f_i, $1 \le i \le m$. Let $(v_1^T\ v_2^T\ \cdots\ v_m^T)^T$ be the corresponding partition of our desired (column) vector v. For v_i, let's try the simplest possible nonzero vector: just one nonzero entry—a 1, say—which should be the first entry

to maximize linear independence among $\{A_i^r v_i\}_{r \geq 0}$. This simple scheme would work if only we had a guarantee that the resulting n-square matrix $B :=$ $(v \ Tv \ \cdots \ T^{n-1}v)$ has linearly independent columns—equivalently, $\det B \neq 0$—based solely on the premise of pairwise relative primeness of the polynomials f_i from which T is formed. But for polynomials, relative primeness means no common root (in an algebraic closure of F) and so a simple numerical quantity guaranteed to be nonzero precisely when the f_i are pairwise relatively prime is the product of all differences of roots of different f_i's. Call this product Δ. We are thus emboldened to conjecture that Δ and $\det B$ are closely related. In fact, as we show in Theorem 2, they are equal (up to sign).

To state this result in appropriate generality, let λ_{ij}, $1 \leq i \leq m$, $1 \leq j \leq n_i$ be algebraically independent commuting indeterminates over F (the "roots") and for $1 \leq i \leq m$ let σ_{ij}, $1 \leq j \leq n_i$, denote the elementary symmetric polynomials in $\{\lambda_{ij}\}_{j=1}^{n_i}$ (the "coefficients"). Then let $f_i(x) = x^{n_i} - \sigma_{i1}x^{n_i-1} + \sigma_{i2}x^{n_i-2} - \cdots + (-1)^{n_i}\sigma_{in_i}$, $1 \leq i \leq m$, so that the roots of f_i are $\{\lambda_{ij}\}_{j=1}^{n_i}$. Set $n = \sum_{i=1}^m n_i$. The companion matrix of f_i is

$$A_i = \begin{pmatrix} 0 & & & & (-1)^{n_i-1}\sigma_{in_i} \\ 1 & 0 & & \mathbf{0} & \vdots \\ & 1 & \ddots & & \\ & & \ddots & 0 & -\sigma_{i2} \\ & \mathbf{0} & & 1 & \sigma_{i1} \end{pmatrix}$$

and the characteristic polynomial of A_i is f_i. Now let A be the n-square block diagonal matrix with A_i as its ith diagonal block, $1 \leq i \leq m$, let e_i denote the n_i-length (column) vector with 1 in the first position and 0's elsewhere, and let e be the n-vector obtained by concatenating the e_i. Finally, let B denote the n-square matrix whose jth column is $A^{j-1}e$, $1 \leq j \leq n$. For example, when $m = 3$ and $n_1 = 3$, $n_2 = 2$, $n_3 = 1$,

$$B = \begin{pmatrix} 1 & 0 & 0 & \sigma_{13} & \sigma_{11}\sigma_{13} & \sigma_{11}^2\sigma_{13} - \sigma_{12}\sigma_{13} \\ 0 & 1 & 0 & -\sigma_{12} & -\sigma_{11}\sigma_{12} + \sigma_{13} & -\sigma_{11}^2\sigma_{12} + \sigma_{11}\sigma_{13} + \sigma_{12}^2 \\ 0 & 0 & 1 & \sigma_{11} & \sigma_{11}^2 - \sigma_{12} & \sigma_{11}^3 - 2\sigma_{11}\sigma_{12} + \sigma_{13} \\ 1 & 0 & -\sigma_{22} & -\sigma_{21}\sigma_{22} & -\sigma_{21}^2\sigma_{22} + \sigma_{22}^2 & -\sigma_{21}^3\sigma_{22} + 2\sigma_{21}\sigma_{22}^2 \\ 0 & 1 & \sigma_{21} & \sigma_{21}^2 - \sigma_{22} & \sigma_{21}^3 - 2\sigma_{21}\sigma_{22} & \sigma_{21}^4 - 3\sigma_{21}^2\sigma_{22} + \sigma_{22}^2 \\ 1 & \sigma_{31} & \sigma_{31}^2 & \sigma_{31}^3 & \sigma_{31}^4 & \sigma_{31}^5 \end{pmatrix} \tag{1}$$

Theorem 2. *For the matrix B formed from the λ_{ij} through their symmetric functions σ_{ij} as in* (1),

$$\det B = \prod_{1 \leq i < j \leq m} \prod_{k=1}^{n_i} \prod_{l=1}^{n_j} (\lambda_{jl} - \lambda_{ik}) \tag{2}$$

Before proving Theorem 2 (and thereby also Kaplansky's Lemma) we make some remarks. The classical resultant of two polynomials expresses the right side of (2) (for $m = 2$) as the determinant of the *Sylvester matrix* [7, p. 102] also formed from the polynomials' coefficients. Thus Theorem 2 gives a form of generalized resultant, though the matrix B does not coincide with the Sylvester matrix for $m = 2$. Further information about resultants (including a generalization in a

different direction) and their application in finding greatest common divisors of polynomials is given in [**2**, §2.1].

What does the matrix B look like when expressed in terms of the λ's? The following two propositions give some useful information.

Proposition 1. *All nonzero terms in the determinant of* any *submatrix of B have the same total degree in the λ's (depending only on the submatrix).*

Proof: The matrix B contains various size identity matrices flush left; the other entries are polynomials in the σ's that are easily seen to be homogeneous polynomials in the λ's whose (total) degrees are as illustrated in Figure 1 (for $m = 3$; $n_1 = 3$, $n_2 = 2$, $n_3 = 1$). Here the blanks represent zero entries in B.

$$\begin{pmatrix} 0 & — & — & 3 & 4 & 5 \\ — & 0 & — & 2 & 3 & 4 \\ — & — & 0 & 1 & 2 & 3 \\ 0 & — & 2 & 3 & 4 & 5 \\ — & 0 & 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ -1 & 0 & 1 & 2 & 3 & 4 \\ -2 & -1 & 0 & 1 & 2 & 3 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ -1 & 0 & 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

Figure 1 **Figure 2**

It is possible to fill in the blanks so that each row is a sequence of consecutive integers, yielding a matrix C as in Figure 2. The degree of the product of given nonzero entries in B is the sum of the corresponding entries in C. Thus it certainly suffices to show: for any square submatrix D of C, the sum of the entries in a diagonal is constant over all diagonals of D (a diagonal is a maximal set of entries, no two in the same row or column). Now the entries in C are of the form $c_{ij} = d_i + e_j$ for some sequences $\{d_i\}, \{e_j\}$ (e_j may be taken $= j$). Thus the sum of the entries in any diagonal of D is the sum of the d_i corresponding to the rows of D and the e_j corresponding to the columns of D, independent of the diagonal. ∎

Proposition 2. *Suppose B is partitioned into m^2 submatrices B_{ij} of size n_i-by-n_j in the obvious way. Fix i in $[1, m]$. Then*

(i) $B_{ii} = A_i^{n_1+n_2+\cdots+n_{i-1}}$,
(ii) $\det B_{ii} = (\lambda_{i1}\lambda_{i2}\cdots\lambda_{in_i})^{n_1+n_2+\cdots+n_{i-1}}$,
(iii) *The degree (in the λ's) of each term in* $\det B_{ii}$ *is* $n_i(n_1 + n_2 + \cdots + n_{i-1})$,
(iv) *If B_i' is any n_i-square submatrix of B involving the same rows as B_{ii} but no column to the right of the rightmost column of B_{ii}, then the degree of each term in* $\det B_i'$ *is strictly less than $n_i(n_1 + n_2 + \cdots + n_{i-1})$ unless $B_i' = B_{ii}$,*
(v) *The degree of each term in* $\det B$ *is* $\Sigma_{1 \le i < j \le m} n_i n_j$.

Proof:

(i) If an n_i-square window slides from left to right along the ith row of the partitioned B, the successive powers $I, A_i, A_i^2, \ldots$ are obtained.
(ii) This is clear since $\det A_i = \sigma_{in_i}$.
(iii) follows from (ii) and Proposition 1.
(iv) follows from (iii) and Proposition 1 since degrees are strictly increasing along each row of B.
(v) also follows from (iii) and Proposition 1. ∎

Now to the proof of Theorem 2 in three steps: a) the right side of (2), Δ divides $\det B$, b) Δ and $\det B$ have the same degree and hence a) implies they can differ

only by a scalar factor, c) Δ and $\det B$ contain an identical term and hence this scalar factor is 1. At this point only step a) requires some work since the degree of Δ is $\sum_{1 \le i < j \le m} n_i n_j$, as is the degree of $\det B$ by Proposition 2(v). Also, the product of the first terms of the factors comprising Δ is $X := (\lambda_{21} \cdots \lambda_{2n_2})^{n_1} (\lambda_{31} \cdots \lambda_{3n_3})^{n_1+n_2} \cdots (\lambda_{m1} \cdots \lambda_{mn_m})^{n_1+n_2+\cdots+n_{m-1}}$ (and this monomial does not occur elsewhere in Δ). On the other hand, by Proposition 2 (iv) and a Laplace expansion of $\det B$, only the entries in the main diagonal blocks $B_{11}, B_{22}, \ldots, B_{mm}$ can produce the monomial X and their contribution is precisely $\det B_{11} \det B_{22} \cdots \det B_{mm} = X$ by Proposition 2 (ii).

It remains to establish a). Clearly, $\det B$ lies in the polynomial ring $F[\lambda_{11}, \ldots, \lambda_{mn_m}]$. First, we show by a standard technique that each factor in Δ is a divisor of $\det B$ in this ring. To do so, suppose $\lambda_{ik} = \lambda_{jl}$ for some i, j, k, l with $i \neq j$. This means that A_i and A_j share a common eigenvalue $\lambda_{ik} = \lambda_{jl} \equiv \lambda$. Let u_i and u_j be *left* (row) eigenvectors of A_i and A_j, respectively, corresponding to λ. Then for all $r \geq 0$,

$$(u_i \, u_j) \begin{pmatrix} A_i & 0 \\ 0 & A_j \end{pmatrix}^r = \lambda^r (u_i \, u_j).$$

We claim the eigenvector u_i can be scaled so that the inner product

$$(u_i \, u_j) \begin{pmatrix} e_i \\ e_j \end{pmatrix} = 0.$$

Since only the first components of e_i, e_j are nonzero, it clearly suffices to show that both inner products $u_i e_i$ and $u_j e_j$ are nonzero. But if, say, $u_i e_i = 0$ then for all $r \geq 0$, $u_i A_i^r e_i = \lambda^r u_i e_i = 0$ while $\{A_i^r e_i\}_{r \geq 0}$ spans all of F^{n_i} (its first n_i vectors are the standard basis for F^{n_i}). Thus $u_i \in F^{n_i}$ would be perpendicular to F^{n_i}, hence 0, and eigenvectors are not allowed to be 0. Thus the claim is established. It follows that

$$(u_i \, u_j) \begin{pmatrix} A_i & 0 \\ 0 & A_j \end{pmatrix}^r \begin{pmatrix} e_i \\ e_j \end{pmatrix} = 0 \qquad \text{for all } r \geq 0.$$

In other words, $(u_i \, u_j)$ provides a nontrivial, yet vanishing, linear combination of the rows of B indexed by its i and j blocks. Thus B is singular and $\det B = 0$. Hence $\lambda_{j\ell} - \lambda_{ik}$ divides $\det B$ in the ring $F[\lambda_{11}, \ldots, \lambda_{mn_m}]$. Since each $\lambda_{jl} - \lambda_{ik}$ is prime in this ring, their product Δ also divides $\det B$. Thus Theorem 2 holds and our generalized resultant yields yet another proof of Kaplansky's Lemma. ■

REFERENCES

1. B. Aupetit, An improvement of Kaplansky's lemma on locally algebraic operators, *Studia Math.* 88 (1988), 275–278.
2. S. Barnett, *Matrices in Control Theory*, Robert E. Krieger Publ. Co., Malabar, Florida, 1984.
3. I. N. Herstein, *Topics in Algebra*, Blaisdell, Waltham, Massachusetts, 1964.
4. N. Jacobson, *Basic Algebra I*, W. H. Freeman, New York, 1985.
5. V. V. Prasolov, *Problems and Theorems in Linear Algebra*, Translations of Mathematical Monographs, Vol. 134, American Mathematical Society, Providence, Rhode Island, 1994.
6. H. Radjavi and P. Rosenthal, *Invariant Subspaces*, Ergeb. Math. Grenzgeb. 77, Springer, Berlin, 1973.
7. B. L. van der Waerden, *Algebra*, Springer-Verlag, New York, 1991.

Department of Statistics
University of Wisconsin-Madison
1210 W. Dayton Street
Madison, WI 53706-1693
callan@stat.wisc.edu

Metric Spaces in Which All Triangles Are Degenerate

Bettina Richmond and Thomas Richmond

In any subspace of the real line **R** with the usual Euclidean metric $d(x, y) = |x - y|$, every triangle is degenerate. In $\mathbf{R}^2$ or $\mathbf{R}^3$ with the usual Euclidean metrics, a triangle is degenerate if and only if its vertices are collinear. With our intuition of a degenerate triangle having "collinear vertices" extended to arbitrary metric spaces, we might expect that a metric space in which every triangle is degenerate must be "linear". It might be reasonable to expect that any "linear" metric space is isometric to a subset of **R** with the usual metric. When classifying all metric spaces that have only degenerate triangles, we find that there are such metric spaces other than (isometric images of) subspaces of **R**. These other spaces, however, all have precisely four points and all are of the same form. In the final section, we illustrate that the usual topology on $\mathbf{R}^2$ can not be generated by any metric in which all triangles are degenerate.

A metric space (M, ρ) is a set of points M with a metric, or distance function, ρ: $M \times M \to [0, \infty)$ that satisfies some natural properties we expect of distances:

$$\rho(x, y) = \rho(y, x) \quad \text{for any } x, y \in M,$$

$$\rho(x, y) = 0 \quad \text{if and only if } x = y, \quad \text{and}$$

$$\rho(x, y) + \rho(y, z) \geq \rho(x, z) \quad \text{for any } x, y, z \in M \text{ (triangle inequality)}.$$

If (M, ρ) and (N, δ) are two metric spaces, a function f: $M \to N$ such that $\rho(x, y) = \delta(f(x), f(y))$ for any $x, y \in M$ is called an *isometry*. Isometries are always one-to-one. The metric spaces (M, ρ) and (N, δ) are *isometric* if there exists an isometry from M onto N. Generally, one does not distinguish between isometric metric spaces.

If (M, ρ) is a given metric space, we say that $\{x_1, x_2, x_3\} \subseteq M$ forms a *degenerate triangle* if $\rho(x_i, x_j) + \rho(x_j, x_k) = \rho(x_i, x_k)$ for some permutation i, j, k of the indices $\{1, 2, 3\}$. Thus, $\{x_1, x_2, x_3\}$ forms a degenerate triangle if the triangle inequality is an equality for some permutation of the points x_1, x_2, and x_3.

We want to study metric spaces in which every triangle is degenerate, that is, metric spaces such that every 3-element subset forms a degenerate triangle. For brevity, such a metric space is called a *degenerate space*, and its metric is called a *degenerate metric*. Triangles with fewer than three distinct vertices are clearly degenerate, and in what follows, we assume that all triangles have three distinct vertices, unless otherwise noted. With this understanding, we can make statements such as "A four-point space has exactly four distinct triangles." The real line with the usual Euclidean metric $d(x, y) = |x - y|$ is denoted by **R**. It is a familiar fact that **R** is a degenerate space. Any subspace of a degenerate space is a degenerate space. Given any degenerate space M, one might suspect that it must be isomorphic to a subspace of **R**, and might attempt to construct an isometry from M into **R**. As we will see, the construction of such an isometry requires that if $\{y, o, p\}$ and $\{o, p, x\}$ are (degenerate) triangles in M with longest sides $\{y, p\}$ and $\{o, x\}$ respectively, then triangle $\{y, o, x\}$ must have longest side $\{y, x\}$. Surprisingly, this

need not be true. Thus, the structure of a degenerate space is based upon the structure of its 4-point subspaces. We consider 4-point degenerate spaces in the next section; in Section 2 we find that the problematic 4-point subspaces cannot occur if the space has more than 4 points.

1. FOUR-POINT DEGENERATE SPACES ARE OF TWO FORMS. Throughout this section, $X = \{S, T, U, V\}$ denotes a 4-point degenerate space. The four points of X are represented by the vertices and center of a (Euclidean) equilateral triangle as shown in Figure 1. The distances between points are denoted by labels on the corresponding edges of this graph. Often we do not distinguish between an edge and its length. By the longest edge of some set, we mean any edge of maximal length from that set. After a permutation of the vertices of X, we may assume that the edge $\{V,T\}$ is maximal among the six edges, and that $\{V, S\}$ is maximal among the remaining four edges having V or T as a vertex. If we let $\rho(V, S) = a$, $\rho(S, T) = b$, $\rho(U, V) = c$, and $\rho(U, T) = d$, then since $\{V, T\}$ is the longest edge in the triangles $\{V, T, S\}$ and $\{V, T, U\}$, we must have $\rho(V, T) = a + b = c + d$, as shown in Figure 2. By the choice of the edge $\{V, S\}$, we know that $a \geq c$. Hence either $\{V, S\}$ or $\{S, U\}$ is the longest edge of the triangle $\{V, S, U\}$, and thus either $\rho(S, U) = a - c$ (Case 1) or $\rho(S, U) = a + c$ (Case 2). Depending on which edge of triangle $\{S, T, U\}$ is the longest, either $\rho(S, U) = b + d$ (Case A), $\rho(S, U) = b - d$ (Case B), or $\rho(S, U) = d - b$ (Case C). Among the 6 cases that result from considering which are the longest edges of these two triangles, the cases 1A, 1B, 2B, and 2C are easily seen to be impossible. For example, in case 1B, the equations $a - c = b - d$ and $a + b = c + d$ yield $a - c = 0$, contrary to S and U being distinct points with $\rho(S, U) = a - c$. In the case 1C, the equations $a - c = d - b$ and $a + b = c + d$ are equivalent, while in the case 2A, the equations $a + c = b + d$ and $a + b = c + d$ yield $a = d$ and $b = c$. The two configurations corresponding to these two cases 1C and 2A are pictured in Figure 3. Both of these are possible, and any 4-point degenerate space must be of one of these two forms.

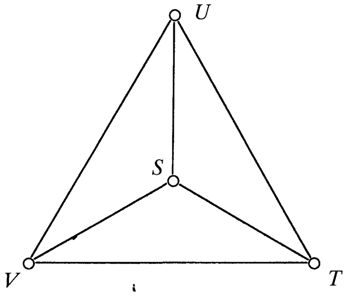


Figure 1

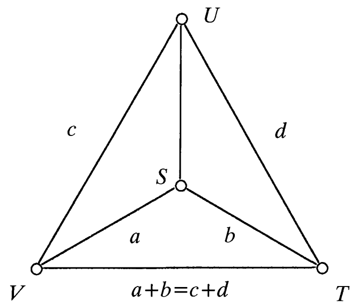


Figure 2

There are familiar models for both of these 4-point metric spaces. The form from Figure 3a can be achieved using Euclidean distance on the real line. If positive distances a, b, and $c < a$ have been chosen, then letting V be any real number, $T = V + a + b$, $S = V + a$, and $U = V + c$ gives a 4-point subspace of $\mathbf{R}$ in which the Euclidean distances between the points agree with those in the form. Because of this model, we call this configuration, shown in Figure 4, the *linear form* of a 4-point degenerate space. The other 4-point degenerate space, shown in

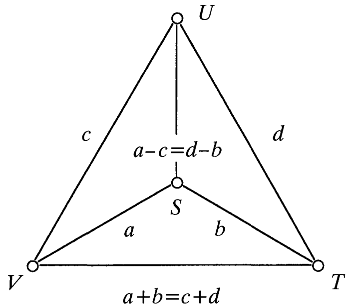


Figure 3a

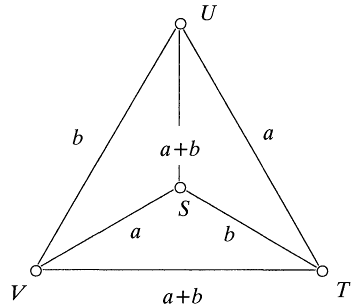


Figure 3b

Figure 3

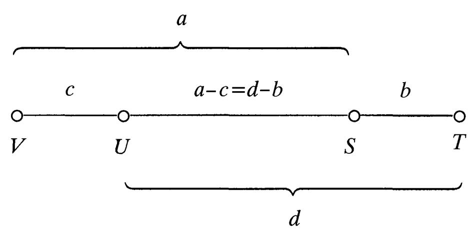


Figure 4. Linear Form.

Figure 3b can be redrawn as a square with its diagonals, as shown in Figure 5a. It should be clear that no such configuration can be realized with positive Euclidean distances, but the square suggests a familiar model, shown in Figure 5b. Let $\{S, U\}$ and $\{T, V\}$ be distinct pairs of diametrically opposite points on a circle with circumference $2a + 2b$ such that the points V and S determine a central angle of $\pi a/(a + b)$ radians. Let the distance between two points on the circle be the length of the shortest arc of the circle connecting the two points. This familiar arc-length metric restricted to the set $\{S, T, U, V\}$ is a realization of the 4-point degenerate space of Figure 3b. Because of this model, we call this configuration the *circular form* of a 4-point degenerate space.

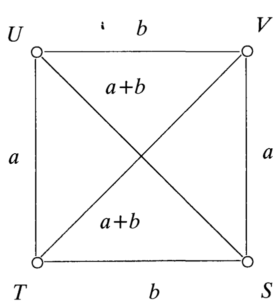


Figure 5a

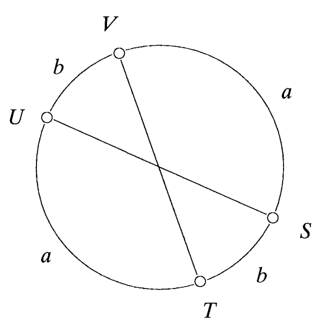


Figure 5b

Figure 5. Circular Form.

2. FIVE-POINT DEGENERATE SPACES ARE OF LINEAR FORM. The structure of any degenerate space is based on its 4-point degenerate subspaces, which must be of linear or circular form. In this section, we show that in any 5-point degenerate space all 4-point subspaces are of linear form. Thus, the problem encountered in defining an isometry from any degenerate metric space M into $\mathbf{R}$ can only arise in 4-point spaces (of circular form).

Before considering general 5-point degenerate spaces, we make a few observations on the specific metrics presented in Section 1 as models of 4-point degenerate spaces. It is clear that extending our Euclidean model of a 4-point degenerate space of linear form to any larger subset of $\mathbf{R}$ results in a degenerate space. However, a degenerate space $\{S, T, U, V\}$ of circular form cannot be extended to any larger degenerate space consisting of points on a circle with the arc-length metric. If R were another point on the circle, relabel the points of $\{S, T, U, V\}$ so that $R, S, T, U,$ and V appear in that order along the circle (traced in any specified direction). Now $R, T,$ and U do not lie in any semicircle, and thus, the sum of the distances from any one of these points to the other two is strictly greater than the arc-length of a semicircle. In our "length of shortest arc" metric, however, no distance can exceed the length of a semicircle. Thus, the triangle formed by $R, T,$ and U cannot be degenerate.

We now show that there is no extension of the 4-point degenerate space of circular form to any 5-point degenerate metric space. We start with the 4-point circular form $\{S, T, U, V\}$ in its square representation and add a fifth point R, as shown in Figure 6, and assume that this space is degenerate. Two opposite edges of the square have lengths x and the other pair of opposite edges of the square have lengths y. Let a be the maximal length of the four edges having R as a vertex. Without loss of generality, we may label the 5-point degenerate space as shown in Figure 7.

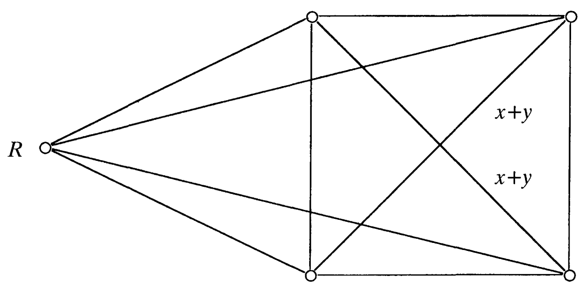


Figure 6

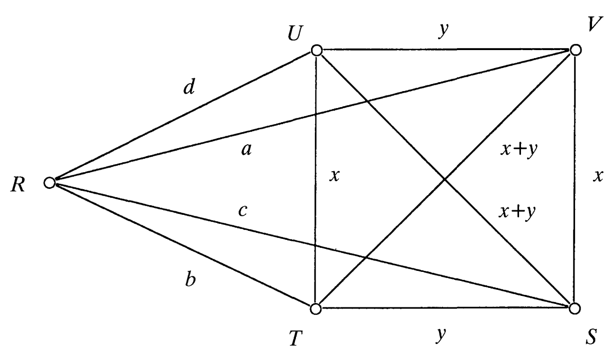


Figure 7

We first show that $a \leq x + y$, that is, $x + y$ is the longest edge of any in Figure 7. Suppose to the contrary that $a > x + y$. Let j be the largest integer such that $a > x + jy$ and let k be the largest integer such that $a > kx + y$. Since a is the length of the longest edge in triangles $\{R, S, V\}$ and $\{R, U, V\}$, we have $a = x + c = d + y$, and thus $c > jy$ and $d > kx$. Adding these inequalities gives $c + d > x + y$, so the longest edge in triangle $\{R, S, U\}$ has length c or d. If c were the longest, then $c = d + x + y > (k + 1)x + y \geq a$, contrary to a being the longest edge incident on R. Similarly, d cannot be the length of the longest edge of $\{R, S, U\}$. Thus, $a \leq x + y$.

Now we show that the assumption that Figure 7 represents a degenerate space leads to the contradiction that the set of real numbers $\{x, y, a\}$ has no maximum. By the preceding paragraph, the longest edges of triangles $\{R, S, U\}$ and $\{R, T, V\}$ have length $x + y$, and thus $a + b = c + d = x + y$. Suppose $x = \max\{x, y, a\}$. The choice of a implies that $x = \max\{x, y, a, b, c, d\}$. From triangles $\{R, S, V\}$ and $\{R, T, U\}$, we have $x = a + c = d + b$. Since $a + b = c + d$, it follows that $a = d$ and $b = c$. Thus, $x = a + c = a + b = x + y$, giving the contradiction that $y = 0$. By the symmetry of x and y, the case $y = \max\{x, y, a\}$ is also impossible. Finally, suppose $a = \max\{x, y, a\}$. From triangles $\{R, V, S\}$ and $\{R, V, U\}$, we have $a = c + x = d + y$. It follows that $2a = c + d + x + y = 2(a + b)$, contrary to $b \neq 0$.

Thus, the 4-point degenerate space of circular form cannot be extended to a 5-point degenerate space, and therefore cannot be extended to any degenerate space with more than 4 points.

3. CLASSIFICATION OF ALL DEGENERATE SPACES. Eliminating circular form subsets from spaces with 5 or more points clears the way for the classification of not only 5-point degenerate spaces, but of all degenerate spaces.

It is clear that every metric space with fewer than three points is degenerate and isometric to a subspace of $\mathbf{R}$. Any degenerate 3-point space consists of one degenerate triangle with edges of length, say, a, b, and $a + b$. Such a space is clearly isometric to the subspace $\{-a, 0, b\}$ of $\mathbf{R}$. Four-point degenerate spaces are of one of the two forms previously described. In any degenerate space with more than four points, every 4-point subspace is degenerate, of linear form.

Suppose M is a degenerate space with more than four points. We will construct an isometry between M and a subspace of $\mathbf{R}$. Pick any two distinct points o and p of M. The point o plays the roll of the "origin" and the point p plays the roll of a "positive point". Define a function f: $M \to \mathbf{R}$ by

$$f(x) = \begin{cases} -\rho(o, x) & \text{if } \{x, p\} \text{ is the longest edge of triangle } \{x, o, p\} \\ \rho(o, x) & \text{if } \{x, o\} \text{ or } \{o, p\} \text{ is the longest edge of triangle } \{x, o, p\}. \end{cases}$$

We first show that f is well defined. Observe that $\{x, p\}$ and $\{x, o\}$ cannot both be longest edges of triangle $\{x, o, p\}$, for then $\rho(x, p) = \rho(x, o) + \rho(o, p)$ and $\rho(x, o) = \rho(x, p) + \rho(p, o)$ lead to the contradiction that $\rho(o, p) = 0$. If $\{x, p\}$ and $\{o, p\}$ are both longest edges of triangle $\{x, o, p\}$, then $\rho(x, p) = \rho(x, o) + \rho(o, p)$ and $\rho(o, p) = \rho(o, x) + \rho(x, p)$, and it follows that $\rho(o, x) = 0 = -\rho(o, x)$, so that both definitions of the value of $f(x)$ agree.

We now show that f is an isometry. If $x = y$, then clearly $\rho(x, y) = 0 = |f(x) - f(y)| = d(f(x), f(y))$. If one of x or y is o, say $y = o$, then $f(x) = \pm\rho(o, x)$, so $\rho(x, o) = |f(x) - 0| = |f(x) - f(o)| = d(f(x), f(y))$. We may thus assume that x, y, and o are distinct. If $\{x, y, o, p\}$ is not already a 4-point subset of M it may be extended to a 4-point subset of M, which, after our unexpected

combinatorial detour, we now know must be of linear form as shown in Figure 4. The linear form of Figure 4 can be linearly ordered in two natural ways: left to right or right to left. Give $\{x, y, o, p\}$ the natural linear order from its linear representation as in Figure 4 in which $o < p$. Without loss of generality, let us assume $x < y$ in this order. There are three cases to consider: either $x < y < o$, $x < o < y$, or $o < x < y$. If $x < y < o$, then $\rho(x, o) = \rho(x, y) + \rho(y, o)$, so $\rho(x, y) = \rho(x, o) - \rho(y, o)$. Since $x < o < p$, $f(x) = -\rho(o, x)$, and similarly $f(y) = -\rho(o, y)$. Thus, $\rho(x, y) = f(y) - f(x)$. Since the distance from x to y is positive, we have $\rho(x, y) = f(y) - f(x) = |f(x) - f(y)|$. If $x < o < y$, then $\rho(x, y) = \rho(x, o) + \rho(o, y)$. Since $o < y$, either $\{o, p\}$ or $\{o, y\}$ is the longest edge of $\{y, o, p\}$, so $f(y) = \rho(o, y)$. As before, $x < o < p$ implies $f(x) = -\rho(o, x)$. Thus, $\rho(x, y) = f(y) - f(x) = |f(x) - f(y)|$. Finally, if $o < x < y$, then $\rho(x, y) = \rho(o, y) - \rho(o, x) = f(y) - f(x) = |f(x) - f(y)|$. Thus, for any choice of $x, y \in M$, we have shown that $\rho(x, y) = |f(x) - f(y)| = d(f(x), f(y))$. This completes the proof that any degenerate space with more than four points is isometric to a subspace of **R**.

Note that if M is a 4-point degenerate space of circular form, the function f defined in the preceding paragraph is still well defined but is not an isometry.

4. TOPOLOGICAL CONSIDERATIONS. Our motivation for considering degenerate spaces originated with a topological question. Denote the topology generated by the Euclidean metric on $\mathbf{R}^n$ by τ_n. With the usual Euclidean metrics, $\mathbf{R}^2$ has non-degenerate triangles but **R** does not. Does this alone provide another verification that $(\mathbf{R}, \tau_1)$ and $(\mathbf{R}^2, \tau_2)$ are not homeomorphic?

There are two immediate observations we should make. First, metrizable topological spaces can be generated by several different metrics which may not share common properties such as boundedness or "degenerateness". In particular, observe that though $(\mathbf{R}, \tau_1)$ is generated by a degenerate metric $d(x, y) = |x - y|$, it is also generated by non-degenerate metrics. If we embed **R** into $\mathbf{R}^2$ as the graph of $y = x^2$ and give this parabola $P = \{(x, y) \in \mathbf{R}^2 : y = x^2\}$ the usual Euclidean distances from $\mathbf{R}^2$, then P has no degenerate triangles, yet P is homeomorphic to $(\mathbf{R}, \tau_1)$. Secondly, it is easily seen that "metrizable with a degenerate metric" is a topological property, for if (M, ρ) is a degenerate space homeomorphic to a topological space X, then a degenerate metric can be defined on X by making the homeomorphism from X to M an isometry. Thus, though the Euclidean metric on **R** is degenerate and the Euclidean metric on $\mathbf{R}^2$ is not, this is not sufficient to conclude that $(\mathbf{R}, \tau_1)$ is not homeomorphic to $(\mathbf{R}^2, \tau_2)$. For this conclusion, we must show that $(\mathbf{R}^2, \tau_2)$ admits no degenerate metric.

If our goal were to show only that $(\mathbf{R}^2, \tau_2)$ is not generated by any degenerate metric, then we could simply note that this is an immediate consequence of the fact that (R^2, τ_2) is not homeomorphic to a subspace of $(\mathbf{R}, \tau_1)$ and the classification of infinite degenerate spaces. However, since we want to use the fact that $(\mathbf{R}^2, \tau_2)$ admits no degenerate metric to show that $(\mathbf{R}^2, \tau_2)$ is not homeomorphic to $(\mathbf{R}, \tau_1)$, we now outline a proof not dependent upon this latter fact.

Suppose ρ is a degenerate metric on $\mathbf{R}^2$ that generates the Euclidean topology τ_2. Pick distinct points a, b, and x from $\mathbf{R}^2$ such that $\{a, b\}$ is the longest side of triangle $\{a, x, b\}$. Thus $\rho(a, x) < \rho(a, b)$ and $\rho(x, b) < \rho(a, b)$. Using the fact that $\rho(a, x)$ is a continuous function of x, if the "middle point" x is moved slightly to $x(t)$, $\rho(a, b)$ remains the longest side of the resulting triangle $\{a, x(t), b\}$. These perturbations of the middle point are actually not restricted to slight movements: in fact, every point z of $\mathbf{R}^2 \setminus \{a, b\}$ is a middle point of $\{a, z, b\}$! Suppose there

exists some point z of $\mathbf{R}^2 \setminus \{a, b\}$ such that $\{a, b\}$ is not the longest side of $\{a, z, b\}$. Let the points $x(t)$ slide along a path in $\mathbf{R}^2 \setminus \{a, b\}$ from x to z. There must exist a first point y on the path for which $\rho(a, y) = \rho(a, b)$ or $\rho(y, b) = \rho(a, b)$. Since $\rho(a, x(t)) + \rho(x(t), b) = \rho(a, b)$ for all the points $x(t)$ on the path before y, it follows that $\rho(a, y) + \rho(y, b) = \rho(a, b)$, and hence $y \in \{a, b\}$. This contradicts the fact that the path avoids the points a and b. Thus, if we select any triangle $\{a, x, b\}$ with x "between" a and b (that is, with longest side $\{a, b\}$), then every point of $\mathbf{R}^2 \setminus \{a, b\}$ must be "between" a and b. In particular, $\rho(a, y) \leq \rho(a, b)$ for every $y \in \mathbf{R}^2$. This implies that our degenerate metric on $\mathbf{R}^2$ is bounded by $2\rho(a, b)$ since $\rho(x, y) \leq \rho(x, a) + \rho(a, y) \leq \rho(a, b) + \rho(a, b)$ for any x and y. Our contradiction will come from the fact that this should hold for any points a and b that happen to form the longest side of some triangle. Pick any two distinct points $z, w \in \mathbf{R}^2$ and let $\epsilon = \rho(z, w)$. Let $\{x_n\}$ be any sequence converging to the origin 0 in $(\mathbf{R}^2, \tau_2)$. Since $\{\rho(x_n, 0)\}$ converges to zero, we may pick points x_j and x_k such that the longest edge of triangle $\{x_j, x_k, 0\}$ has length less than $\frac{\epsilon}{2}$. But ρ is bounded by twice the length of the longest edge of this triangle, contrary to the choice of $\epsilon = \rho(z, w)$ for some $z, w \in \mathbf{R}^2$. This shows that $(\mathbf{R}^2, \tau_2)$ does not admit any degenerate metric, and thus is not homeomorphic to $(\mathbf{R}, \tau_1)$.

Using the topological property "metrizable by a degenerate metric" to distinguish between $\mathbf{R}$ and $\mathbf{R}^2$ is a good exercise in manipulating metrics, but one should recognize the dependence of our proof upon another more common topological property. The argument that if x is "between" a and b, then every $y \in \mathbf{R}^2 \setminus \{a, b\}$ is "between" a and b utilized the fact that $\mathbf{R}^2 \setminus \{a, b\}$ is path connected (relative to τ_2). Removing two points from $(\mathbf{R}, \tau_1)$ never gives a path connected space, providing a more standard argument that $(\mathbf{R}, \tau_1)$ is not homeomorphic to $(\mathbf{R}^2, \tau_2)$. On the other hand, proving that $(\mathbf{R}^2, \tau_2)$ is not metrizable by a degenerate metric shows not only that $(\mathbf{R}^2, \tau_2)$ is not homeomorphic to $(\mathbf{R}, \tau_1)$, but also that $(\mathbf{R}^2, \tau_2)$ is not homeomorphic to any subspace of $(\mathbf{R}, \tau_1)$.

REFERENCES

1. Kaplansky, I., *Set Theory and Metric Spaces*, Chelsea Publishing Co., New York, 1977.
2. Munkres, J. R., *Topology, A First Course*, Prentice-Hall Inc., Englewood Cliffs, NJ, 1975.

Department of Mathematics
Western Kentucky University
1 Big Red Way
Bowling Green, KY 42101
Bettina.Richmond@wku.edu
Tom.Richmond@wku.edu

Partitions of Unity for Countable Covers

Albert Fathi

This paper should be considered as expository classroom notes for the instructor.

Existence of partitions of unity for metric spaces is usually proved via some rather exacting set-theoretical and topological arguments using the (equivalent) concept of paracompactness. Although the standard proof of paracompactness of metric spaces is the one given by M. E. Rudin [**Ru**], in his 1965 PhD thesis, Michael Mather showed that it is easier, for metric spaces, to show directly the existence of locally finite partitions of unity for arbitrary covers [**Ma**]. This proof does not seem to be so well-known although it has been reproduced in [**Bo**] (see also the appendix of [**Do**]). It is astonishing that Mather's arguments do not appear in recent textbooks. We hope that the following exposition for countable covers will popularise it. An advantage of the method is that the same proof can be used in the smooth category.

Although it is formally covered by the countable case, we will first explain the method in the case of finite covers. This will show how easy it is to obtain partitions of unity for compact metric spaces.

For the sake of completeness, let us recall a few definitions. If X is a topological space, the *support* of a continuous function $\varphi: X \to \mathbb{R}$ is the closed set $\operatorname{supp}(\varphi) = \overline{\{x \mid \varphi(x) \neq 0\}}$. A *partition of unity* on X is a family $(\varphi_i)_{i \in I}$ of continuous functions $\varphi_i: X \to [0,1]$ such that $\sum_{i \in I} \varphi_i(x) = 1$ for every $x \in X$. Such a partition of unity is *locally finite* if for every $x \in X$ there exists a neighbourhood V_x of x such that $\{i \in I \mid V_x \cap \operatorname{supp}(\varphi_i) \neq \emptyset\}$ is finite. If $(U_i)_{i \in I}$ is an open cover of X, a partition of unity *subordinated to the cover* $(U_i)_{i \in I}$ is a partition of unity $(\varphi)_{i \in I}$ such that $\operatorname{supp}(\varphi_i) \subset U_i$ for every $i \in I$.

1. FINITE COVERS OF METRIC SPACES. Let $U_1, \dots, U_n$ be an open cover of the metric space X. We wish to construct a partition of unity $(\varphi_i)_{1 \le i \le n}$ such that $\operatorname{supp}(\varphi_i)$ is contained in U_i for each $i = 1, \dots, ,n$.

We start with the continuous functions $f_i(x) = d(x, X \setminus U_i)$ and we set $F(x) = \sum_{i=1}^n f_i(x)$. Observe that $F(x) > 0$ everywhere, since $U_1, \dots, U_n$ is a cover of X and $f_i > 0$ on U_i. Next, we define a continuous function $g_i: X \to [0, \infty[$ by

$$g_i(x) = \max\left(f_i(x) - \frac{1}{n+1}F(x), 0\right).$$

We claim that $\operatorname{supp}(g_i) \subset U_i$ and $\sum_{i=1}^n g_i(x) > 0$ everywhere.

To prove the first assertion, note that $\operatorname{supp}(g_i) = \overline{\{x \mid f_i(x) > F(x)/(n+1)\}}$ is contained in the closed set $\{x \mid f_i(x) \ge F(x)/(n+1)\}$. Since $F(x) > 0$ everywhere, the closed set $\{x \mid f_i(x) \ge F(x)/(n+1)\}$ is itself contained in $U_i = \{x \mid f_i(x) > 0\}$.

To prove the second assertion, write

$$\sum_{i=1}^n g_i(x) \ge \sum_{i=1}^n \left(f_i(x) - \frac{1}{n+1}F(x)\right) = F(x) - \frac{n}{n+1}F(x) = \frac{1}{n+1}F(x).$$

To obtain the partition of unity $(\varphi_i)_{1 \le i \le n}$, define $\varphi_i(x) = g_i(x)/(\sum_{i=1}^n g_i(x))$.

2. COUNTABLE COVERS OF METRIC SPACES. Let X be a metric space and let $(U_i)_{i\in\mathbb{N}}$ be a countable cover. We will modify our argument to show how to construct a locally finite partition of unity $(\varphi_i)_{i\in\mathbb{N}}$ such that $\operatorname{supp}(\varphi_i)$ is contained in U_i for $i = 1, 2, \ldots$.

Define $f_i: X \to [0, 2^{-i}]$ by

$$f_i(x) = \min\left[d(x, X \setminus U_i), 2^{-i}\right].$$

Then $f_i > 0$ on U_i, $f_i = 0$ outside U_i, and

$$\|f_i\|_0 = \sup\{|f_i(x)| \mid x \in X\} \leq 2^{-i}.$$

This inequality implies that the function $F = \sum_{i=0}^{\infty} 2^{-i} f_i$ is continuous; moreover, since $(U_i)_{i\in\mathbb{N}}$ is a cover of X and $f_i > 0$ on U_i, we have $F > 0$ everywhere.

For $i \in \mathbb{N}$, we define the continuous function $g_i: X \to [0,1]$ by

$$g_i(x) = \max\left(f_i(x) - \tfrac{1}{3}F(x), 0\right).$$

We claim that $(g_i)_{i\in\mathbb{N}}$ is a locally finite family of functions such that $\operatorname{supp}(g_i) \subset U_i$ and $\sum_{i=0}^{\infty} g_i(x) > 0$ everywhere. The argument to show $\operatorname{supp}(g_i) \subset U_i$ is almost identical to the one given in the case of a finite cover.

Let us show that $(g_i)_{i\in\mathbb{N}}$ is a locally finite family of functions. Let x be fixed in X. Since F is continuous and positive everywhere, there exists a neighbourhood V of x and an $\epsilon > 0$ such that $F(y) > \epsilon$ for every $y \in V$. Let i_0 be such $2^{-i_0} < \epsilon/3$; from the definition of g_i and the fact that $\|f_i\|_0 \leq 2^{-i}$, it follows that $g_i(y) = 0$ for every $y \in V$ and every $i \geq i_0$.

It remains to show that for every x one can find an i such that $g_i(x) > 0$. Fix an $x \in X$. Since $f_i(x) > 0$ for some i and $\|f_n\|_0 \leq 2^{-n}$, there must be some i_0 such that $f_{i_0}(x) = \sup_{i\in\mathbb{N}} f_i(x) > 0$. From the definition of F, we obtain $F(x) \leq (\sum_{i=0}^{\infty} 2^{-i}) f_{i_0}(x) = 2 f_{i_0}(x)$. It follows that $g_{i_0}(x) \geq f_{i_0}(x) - 2f_{i_0}(x)/3 = f_{i_0}(x)/3 > 0$.

Finally, define the family $(\varphi_i)_{i\in\mathbb{N}}$ by

$$\varphi_i = \frac{g_i}{\sum_{j=0}^{\infty} g_j}, \quad i = 1, 2, \ldots.$$

3. SMOOTH PARTITIONS OF UNITY FOR OPEN SETS OF SEPARABLE HILBERT SPACES. Existence of smooth partitions of unity for an open cover of a manifold has been established in various contexts. The finite dimensional case is well-known; for the infinite dimensional case see **[La]**, **[To]**. We show that the method already given covers the separable case. We will do that for open sets of Hilbert spaces for sake of simplicity.

Let H be a separable Hilbert space (this covers also the finite dimensional spaces $\mathbb{R}^n$). Since the square of the norm $\|\cdot\|$ is a quadratic form, the map $x \mapsto \|x\|^2$ is smooth $(= C^\infty)$ with first derivative bounded on every bounded set; the second derivative is constant and higher order derivatives are 0. Composing this map with an appropriate smooth map with compact support $\mathbb{R} \to [0,1]$, we obtain the following lemma.

Lemma 3.1. *If $x \in H$ and $0 \leq r < s$, there exists a smooth function $\varphi: H \to [0,1]$ such that $\varphi > 0$ on $B(x,r)$, $\varphi = 0$ outside $B(x,s)$ and $\{\sup\|D^k\varphi(x)\| \mid x \in H\} < \infty$ for every $k \in \mathbb{N}$.*

Before stating the second lemma, we recall that a *Fréchet space* is a complete Hausdorff topological vector space whose topology is defined by a countable family of semi-norms.

Lemma 3.2. *Let F be a Fréchet space and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in F. There exists a sequence $(\epsilon_n)_{n\in\mathbb{N}}$ of positive numbers such that for every sequence of scalars $(\lambda_n)_{n\in\mathbb{N}}$ with $|\lambda_n| \leq \epsilon_n$ for all $n \in \mathbb{N}$, the series $\sum_{n\in\mathbb{N}} \lambda_n x_n$ converges in F.*

Proof: Let $(p_n)_{n\in\mathbb{N}}$ be a sequence of semi-norms defining the topology of F. Any sequence $(\epsilon_n)_{n\in\mathbb{N}}$ of positive numbers such that $\epsilon_n \max_{0\leq k\leq n} p_k(x_n) < 1/2^n$ will do. ∎

The space of smooth functions $\varphi: H \to \mathbb{R}$ with uniformly bounded derivatives–i.e., such that $\sup\{\|D^k\varphi(x)\| \,|\, x \in H\} < \infty$ for every $k \in \mathbb{N}$—is a Fréchet space in which $p_k(\varphi) = \sup\{\|D^k\varphi(x)\| \,|\, x \in H\}$ is the family of semi-norms.

Let U be an open set in H and let $(U_i)_{i\in I}$ be an open cover of U. We wish to construct a smooth partition subordinated to that cover. Using the separability of H, we can find a countable family of balls $(B(x_n, r_n))_{n\in\mathbb{N}}$ such that the family $(B(x_n, r_n/2))_{n\in\mathbb{N}}$ covers U and for each n there exists an $i \in I$ such that $B(x_n, r_n) \subset U_i \subset U$. Using Lemma 3.1, let f_n: $H \to [0,1]$ be a smooth function, with all derivatives uniformly bounded, such that $f_n > 0$ on $B(x_n, r_n/2)$ and $f_n = 0$ outside $B(x_n, r_n)$. Multiplying f_n by 2^{-n}, if necessary, we can assume that $\|f_n\|_0 \leq 2^{-n}$.

By Lemma 3.2, we can find a sequence $(\lambda_n)_{n\in\mathbb{N}}$ of positive numbers such that $\lambda = \sum_{n\in\mathbb{N}} \lambda_n < \infty$ and $F = \sum_{n\in\mathbb{N}} \lambda_n f_n$ is a smooth function that is positive everywhere.

Let θ: $\mathbb{R} \to [0,\infty[$ be a smooth function such that $\theta^{-1}(0) =]-\infty, 0]$. If we set

$$g_n(x) = \theta\left[f_n(x) - \frac{1}{\lambda+1}F(x)\right],$$

we can check that the family $(\varphi_n)_{n\in\mathbb{N}}$ defined by

$$\varphi_n = \frac{g_n}{\sum_{k=0}^{\infty} g_k}$$

is a locally finite smooth partition of unity subordinated to the cover $(B(x_n, r_n))_{n\in\mathbb{N}}$.

We can now use the following well-known argument to find a locally finite partition of unity $(\psi_i)_{i\in I}$ subordinated to the cover $(U_i)_{i\in I}$: For each $n \in \mathbb{N}$, choose $i_n \in I$ with $B(x_n, r_n) \subset U_{i_n}$ and set $\psi_i = \sum_{\{n \,|\, i_n = i\}} \varphi_n$. We have thus obtained

Theorem 3.3. *Let U be an open set in a separable Hilbert space. For any countable open cover $(U_i)_{i\in\mathbb{N}}$ of U there is a partition of unity $(\varphi_i)_{i\in\mathbb{N}}$ subordinated to $(U_i)_{i\in\mathbb{N}}$ with φ_i smooth, $i = 1, 2, \ldots$.*

Remark 3.4. If M is a σ-compact smooth manifold of finite dimension, the space of smooth functions on M is a Fréchet space for the C^∞ compact-open topology. We can then use arguments very similar to the ones given here to construct smooth partitions of unity on M.

4. ANOTHER WELL-KNOWN CONSEQUENCE OF LEMMA 3.2. Lemma 3.2 can be used in other instances to give easy proofs of useful facts. We illustrate this with the following proposition.

Proposition. *Let M be a σ-compact smooth manifold of finite dimension. Suppose U is an open set and X is a smooth vector field on U. There exists a smooth function φ: $M \to [0,\infty[$ such that $\varphi > 0$ everywhere on U and $(\varphi|U)X$ can be extended by 0 to a smooth vector field on the whole of M.*

Proof: The space $C^\infty(M, \mathbb{R})$ of smooth functions and the space $\mathscr{T}^\infty(M)$ of smooth vector fields are both Fréchet spaces for the C^∞ compact open topology. Let $(\varphi_n)_{n \in \mathbb{N}}$ be a smooth partition of unity with *compact* supports on U. Since $\operatorname{supp}(\varphi_n)$ is a compact subset of U, we can extend φ_n by 0 to a smooth function on M. In the same way we can extend by 0 the smooth vector $\varphi_n X$ defined on U to a smooth vector field X_n defined on M. We can apply Lemma 3.2 twice, to the sequence φ_n in the Fréchet space $C^\infty(M, \mathbb{R})$ and to the sequence X_n in the Fréchet space $\mathscr{T}^\infty(M)$, to find a sequence λ_n of strictly positive numbers such that $\sum_{n=0}^{\infty} \lambda_n \varphi_n$ converges in $C^\infty(M, \mathbb{R})$ and $\sum_{n=0}^{\infty} \lambda_n X_n$ converges in $\mathscr{T}^\infty(M)$. Since $(\varphi_n)_{n \in \mathbb{N}}$ is a partition of unity on U and $\lambda_n > 0$, the function $\varphi = \sum_{n=0}^{\infty} \lambda_n \varphi_n$ is positive on U. Moreover, the vector field $\sum_{n=0}^{\infty} \lambda_n \varphi_n X_n$ is smooth on the whole of M and, on U, this sum is $\sum_{n=0}^{\infty} \lambda_n \varphi_n X$, which is equal to φX. ■

REFERENCES

[Bo} N. Bourbaki, *Eléments de Mathématiques: Topologie Générale Chapitres* 5 *à* 10, Hermann, Paris, 1974.

[Do] A. Dold, *Lectures on Algebraic Topology*, Grundlehren der Mathematischen Wissenschaften 200, Springer-Verlag, Berlin, 1972.

[La] S. Lang, *Differential and Riemannian manifolds*, Third Edition. Graduate Texts in Mathematics, Springer-Verlag, Berlin, 1994.

[Ma] M. Mather, *Paracompactness and partitions of unity*, Mimeographed Notes 1964; PhD thesis, Cambridge University, 1965.

[Ru] M. H. Rudin, *A new proof that metric spaces are paracompact*, Proc. Amer. Math. Soc., **20** (1969), 603.

[To] H. Torunczyk, *Smooth partitions of unity in some non-separable Banach spaces*, Studia Math. **46** (1973), 43–51.

Unité de Mathématiques Pures et Appliquées
CNRS-UMR 128
Ecole Normale Supérieure de Lyon
46, allée d'Italie
69364 Lyon Cedex 07, FRANCE
afathi@umpa.ens-lyon.fr

From fifty years ago in the MONTHLY...

Results from the Seventh Annual William Lowell Putnam Mathematical Competition held May 24, 1947.

The second prize, three hundred dollars, is awarded to the Department of Mathematics of Yale University, New Haven, Connecticut. The members of the team were Murray Gell-Mann, Murray Gerstenhaber, and Henry Otto Pollak; to each of these a prize of thirty dollars is awarded.

p. 400, vol. 54, 1947

Calculus: A Modern Perspective

Jeff Knisley

1. INTRODUCTION. An American history curriculum that ended with the Civil War would be no more acceptable than a philosophy curriculum that ended with Kant. Yet an acceptable history of mathematics curriculum gives little more than a cursory nod to the mathematics of the twentieth century. This is not to say that the two hundred years following Newton and Leibniz do not deserve seven chapters in a history of mathematics textbook, but rather that the one hundred years leading up to the present deserve more than one [**1**].

Unfortunately, our entire undergraduate curriculum has the same focus as our history of mathematics course—the two hundred years following Newton and Leibniz. Our graduates are more prepared for the period in which the steam engine replaced the horse than they are for the period in which compact disks replaced the vinyl LP. It is no wonder that a reform movement emerged early in this century, as evidenced by several articles in [**2**], nor is it surprising that the reform movement is stronger than ever today. Tragically, many mathematicians have responded to reform as in [**3**], where it was lamented that "Mathematics is losing its soul. Its priests are pawning it off to a different god." Such a call to arms only reinforces the popular image of mathematicians as the last practitioners of some ancient art.

We know better. We know that mathematics is growing and thriving, fruitful and strong. However, the new growth in mathematics is all but absent from our undergraduate curriculum. Indeed, our traditional calculus course is packed with intellectual deadwood—contrived applications, outdated examples, and obsolete definitions. It is time we allowed some new growth in a curriculum that has been antiquated for most of this century [**7**].

2. THE NEW GROWTH. Much of modern mathematics is derived from modern trends in calculus. Many of the ideas in differential geometry, statistics, and numerical analysis are descended from the study of calculus in the present century. Correspondingly, any new ideas introduced into the calculus curriculum should be cultivated from these modern trends. Technology allows us to incorporate linear regression, Markov processes, and probability distributions into an introductory calculus course, and when we do so, our students sense that they are studying ideas relevant to the world in which they live.

Similarly, students appreciate applications that have a modern perspective, and such applications need not be outside of mathematics [**4**]. Applications of the derivative can be enhanced with the study of cubic splines and Bezier curves. The study of Newton's method can be generalized to the study of fixed point theorems in general. The study of spectral theory begins with boundary value problems such as

$$y'' = -\alpha^2 y \tag{1}$$

$$y(0) = y(\pi) = 0 \tag{2}$$

and students genuinely enjoy being shown that (1)–(2) has nontrivial solutions only for integral values of α.

Finally, allowing such new growth can greatly simplify our efforts at instruction. Complex numbers make partial fractions and trigonometric substitutions much more accessible. A simple matrix exponential is a great illustration of the power series concept. And stating the laws of exponents as axioms for an abelian group shows a student that these are more than rules for manipulating superscripts.

3. THE DEAD WOOD. Allowing new growth into the calculus curriculum means something old must go, but such additions mean more than simple pruning. Indeed, patching the new into the old destroys the continuity and coherence of the original structure. Our present curriculum bears witness to this fact. The applications of the integral contribute little to the remainder of the course, and the customary chapter on analytic geometry seems misplaced. Sequences are introduced in the context of convergence of series, and thus it should be no surprise that students get the two confused.

Rather than further fragment our curriculum, we need to transform the calculus so that it is a coherent mix of timeless concepts and new ideas. I believe such a transformation must address the following:

Approximation and limits. Our current calculus course relies on several seemingly unrelated notions of limit. The $\epsilon - \delta$ definition is introduced en route to the definition of the tangent line. Infinite limits are defined using ϵ and N sufficiently large. Newton's method is left to intuition. The definite integral is defined using the norm of a partition that goes to zero. Numerical integration introduces the idea of bounded error. Limits of sequences are defined with the monotone convergence theorem. And after spending section after section using limits of sequences to develop convergence tests, the idea of a converging Taylor series is developed from the remainder formula for Taylor polynomials. No attempt is ever made to connect all these ideas of limit into a coherent concept.

We need to introduce and define the limit so that all our applications of approximation and convergence are derived from a single concept. This concept will likely have to be a principle rather than a definition. It may be more appropriate to explore approximation with a graphing calculator than with a formal system of definitions and theorems.

Intuition and Rigor. We prove that the Mean Value Theorem is a consequence of the extreme value theorem, but we do not prove the extreme value theorem itself. Instead, we argue that the extreme value theorem is intuitively obvious. The reason for such an intuitive appeal is that the proof of the extreme value theorem depends on the Heine Borel Theorem, and the Heine Borel Theorem depends on the topology of the real line. Thus, the proof of the Mean Value Theorem is not a proof at all. We might as well argue that the Mean Value Theorem is intuitively obvious and skip its proof altogether.

But we should not skip the Mean Value Theorem altogether. The point is that an introductory calculus course is not about theorems but is rather about definitions. If the difference operator

$$\Delta f(x) = f(x+h) - f(x)$$

could be used easily, then we would not even need the concept of a limit. However, our computationally attractive chain and product rules come solely from the

definition

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Good theorems are the stuff of graduate courses. Good definitions are the stuff of introductory calculus.

In my mind, this all but eliminates the traditional Riemann definition of the integral. It requires too much time and machinery—Riemann sums, the norm of a partition, the arbitrary choice of a point in a subinterval—and it is machinery that will not be used again. In contrast, Lesbegue's definition of the integral is intuitively simple and can be stated without a lot of machinery. The integral of a simple function is a picture. It can be computed using a table. The integral is a measure of simple function approximation. Simple functions can be used later in developing the integral test for convergence of a positive term series.

Technology and Modern Science. The computer was developed by mathematicians like Von Neumann for mathematicians like us, so it is absurd that mathematicians would not enthusiastically embrace their own creation. Technology should be a tool we welcome with excitement—no more rigging problems so the algebra comes out right. If a problem requires the solution to

$$6x^5 + 3x^2 + 1 = 0,$$

then the student can pull out the trusty graphing calculator, estimate the roots graphically and use a root finder to polish off the answer to the desired number of decimals. We should forget extremum problems with functions like

$$f(x) = x^6 + x^3 + x,$$

and instead to ask them to find the extrema of functions like

$$f(x) = \int_0^x \frac{6t^5 + 3t^2 + 1}{\sqrt{t^4 + 1}}\, dt.$$

As a result, calculus can be presented to the student as the true foundation of modern mathematics and science instead of as a hodge-podge of problems restricted to angle multiples of 30° and 45°. Indeed, modern science is more than the study of classical mathematics. It runs parallel to modern mathematics and often intertwines with it. Technology means that regression and curve-fitting can be studied in a calculus course and then immediately applied in a chemistry course. Our debate should be about how we should use technology, not if we should use technology.

4. CONCLUSION. The present dilemma is both fortunate and obvious. Calculus *is* modern, but our calculus curriculum is not. Calculus will continue to prosper with time and technology. We can take comfort in knowing that mathematics will continue to enlighten and enliven the minds of generations to come.

But poetry does not have to be taught by poets, and likewise, mathematics need not be taught by mathematicians. Our traditional course barely even tests the abilities of software tools such as Maple and Mathematica, and the days when every student carries a laptop are not that far off. Already our first semester calculus course is little more than a supplement to the high school curriculum. Without a modern perspective and an enthusiasm for the technology we mathe-

maticians created for ourselves, our relevance to society will continue to dwindle. It will be only a matter of time before we mathematical horsemen are replaced by the intellectual equivalent of the steam engine.

REFERENCES

1. Eves, H., *An Introduction to the History of Mathematics*, 5th edition, Saunders College Publishing, Philadelphia, 1983.
2. Ewing, J., *A Century of Mathematics—Through the Eyes of the Monthly*, The Mathematical Association of America, Washington, 1994.
3. Greenman, C., On Articles in the October Issue, *Notices Amer. Math. Soc.* 43 (1996), 5–6.
4. Kleinfeld, M., Calculus: Reformed or Deformed, *Amer. Math. Monthly* 103 (1996), 230–232.
5. Krantz, S., Math for Sale, *Notices Amer. Math. Soc.* 42 (1995), 1116.
6. Royden, H. L., *Real Analysis*, 2nd edition, MacMillan Publishing Company, Inc., New York, 1968.
7. Woods, R., How Can Interest in Calculus be Increased?, *Amer. Math. Monthly* 36 (1929), 28–32.

Department of Mathematics
East Tennessee State University
Johnson City, TN 37614-0663
knisleyj@etsuarts.etsu-tn.edu

Book review from the MONTHLY fifty years ago...

College Algebra. By A. A. Albert, New York and London, McGraw-Hill Book Co., 1946. 12 + 278 pages. $2.75

In his introduction the author justifies his new textbook in college algebra and one can do no better than to let him present his own case: "College algebra has been a most abused subject. The time allotted to it is frequently inadequate for a genuinely good treatment, and indeed the entire course is sometimes omitted. This is due partly to a desire to bring students to a study of the calculus as early as possible. It is also due partly to the presentation of college algebra, in all texts thus far published, as a collection of seemingly unrelated topics. The desire to teach the calculus as early as possible tends to defeat its own ends. The building of a course in the calculus on what must be a weak foundation cannot result in a good student understanding of the subject. There is also no reason why the material of college algebra cannot be cohesively organized."

p. 174, vol. 54, 1947

Pro Choice

Arnold Ostebee and Paul Zorn

INTRODUCTION. This essay is "pro reform." But before arguing for "reform," a few important words about the word itself. First, we acknowledge some discomfort with the value-laden term "reform"; such judgments are better left to the reader. But the word is now in common use, so we'll adopt it and its variants without further apology, and usually without quotation marks. Second, we observe that there is no such thing as *the* reformed approach to calculus. We self-styled reformers do indeed tend to subscribe to some common broad goals and tenets, such as the primacy of conceptual understanding, the pedagogical value of multiple representations, and the potential of technology to improve both pedagogy and content. On the other hand, there is no party line—let alone unanimity—among reformers on much else. Even a cursory comparison of several reform texts shows that there is no single reformed position on convergence and divergence of numerical series, no reform-approved list of integration techniques, no official ban on lecturing, no accepted dogma on implicit differentiation, and no enforced taboo on the mean value theorem.

Individual reformers and texts do, of course, take positions on most of these questions: a text either covers trigonometric substitutions or it doesn't. That different reformers make different choices—indeed, reformed approaches offer much more diversity of choice than do traditional texts—is no valid criticism of the reform idea in general. On the contrary, the willingness of reformers to make hard choices is an important virtue of reformed approaches to calculus, and the one we praise in this essay.

Facing choice. Teaching calculus, "reformed" or not, will always be a matter of choices: what to cover, what to prove, what to test, what to assign, what to say, what to imply, what to skip, which tools to allow, which to forbid. We may not consciously acknowledge or even recognize these choices, but they are inescapable, and they have consequences. If we study hypergeometric series deeply, we may not get to numerical integration—and vice versa. Is Euler's method a fair trade for partial fractions? Is a proof of the chain rule worth the same amount of time spent understanding what it says? Is the course itself an introduction to mathematical analysis, or is it an introduction to calculus-based tools and their uses?

The key point is simple: the inevitable choices we face—as "reformers" or as "traditionalists"—are very often among *competing goods.* (In the heat of mathematico-political debate, combatants often caricature the choices as between good and evil.) Related rates problems, for instance, figure on many reformers' hit lists, and for good reason: too often, the problems are contrived variations on a few rigid themes. But there's another side to the story. Understanding derivatives as rates is a standard tenet of calculus reform, so the idea of related rates—relating varying quantities to the rates at which they vary—can be seen as a natural and productive extension. In other words, the topic of related rates is intrinsically neither good nor bad; what matters is what we choose to make of the topic, and why.

The *need* to make hard pedagogical choices in calculus has nothing special to do with a reformed approach to the subject. A *willingness* to make hard choices, however, does seem to us to be a special virtue of reformed approaches. As textbook authors we know well (and may sometimes even have succumbed to) the temptation to choose some of everything. With world enough and time that strategy might work, but the real world of calculus is too large, and the time available too limited, to permit that luxury.

How much rigor? As examples of hard choices, consider some questions of rigor: Should important theorems and definitions be stated in full generality? Should they be proved in the same spirit?

A mathematician's natural and (usually) healthy impulse is, when in doubt, to prove everything, as generally as possible. Granted, no modern elementary calculus text follows this policy consistently, but its indirect effects are clear: most traditional texts contain too many proofs. "Reformed" texts tend to be more selective. Some of these proofs are quite subtle, although fine points may be finessed with judicious disclaimers. Even if the proofs elude most students (goes one argument), some of the best will understand, and the rest will at least have "seen" the proofs. These treasured seeds, though little watered now, may someday sprout.

This "broadcast" strategy fails on both mathematical and pedagogical grounds. Careful proofs, even of the "simplest" theorems in calculus, often depend on quite subtle ideas from analysis, such as the persistence of equalities and inequalities upon taking limits. In addition, the lesson many students draw from half-understood, force-fed rigor is damaging and discouraging: We aren't even *supposed* to understand this stuff, and more like it is probably on the way.

Many reformers (ourselves included) choose simply to skip most formal proofs in elementary calculus. To put it more politely, we "defer" such proofs until later courses in analysis. This policy has, at the very least, the virtue of forthrightness—there is no pretense of presenting a logically airtight development of the subject. Just as important is the pedagogical calculation involved. The time and energy even "easy" proofs take is often much better spent helping students understand what a few important theorems really *say*: the difference between hypotheses and conclusions, what goes wrong without a key hypothesis, whether or not the converse holds, how the theorem answers a "natural" question, etc. The goal of understanding theorems, by the way, is much too often taken for granted. If you doubt it, try asking Calculus II students to state—not prove—*any* important theorem from Calculus I. Then stand back and watch the syntax fly.

The concrete approach to rigor we take in our own calculus textbook **[1]**, is to omit *most* proofs, but to present a chosen few (very few) in some detail, as a short but moderately serious excursion into new mathematical terrain. We try to explain, for instance, that while the implication

$$f' = 0 \Rightarrow f = \text{constant}$$

may *seem* obvious, it really requires proof, and the proof is surprisingly subtle. It depends on a version of a mean value principle; we prove one in some detail. The result we aim for, in the short run, is not primarily to settle students' "live" questions of logical validity, but rather to introduce students, briefly but honestly, to an important area of mathematical culture. Later, we hope, many students will revisit the area.

The hardest choices. As textbook authors we and other reformers have had to face many of the hard choices implicit in designing calculus courses; quite properly,

we've faced them in different ways. Authors don't make *every* choice, of course. Some are best made by individual teachers; still other choices, we admit, can be finessed or evaded by dint of appendices, "optional" sections, and other devices of creative publication. Hardest and most important are the highest-level choices, such as those related to a course's general goals, its target audience, and the "toolkit" it employs. These choices reverberate everywhere throughout a course.

Consider, for instance, the molten-button issue of building skill and speed with paper-and-pencil symbol manipulation, and its countless corollary questions: What traditional manipulation techniques are really important? Are some techniques obsolete in the age of the TI-92 and its inevitable successors? How should by-hand and mental calculation be balanced? Is symbolic manipulation facility (by hand or by head) essential for conceptual understanding? Do we learn by (physical) doing? *Must* we learn this way?

Reasonable people, reformers included, can ... and do ... and should ... differ, significantly, on all of these questions. We say "reasonable" with full sincerity: To downplay the algebraic viewpoint, even radically, in the limited agenda of an elementary calculus course is not to deny its importance in the larger mathematical sphere. By the same token, there is nothing inherently "mindless" about symbolic drill, any more than with any other form of mental or physical exercise. The key, at either extreme or anywhere in between, is a clear and balanced view of a course's larger goals and intended audience.

Conclusion. Calculus reform is not a single alternative to a standard diet; reform offers a diverse menu of different but carefully-considered choices. The greatest lasting value of calculus reform may well be in highlighting and forcing important choices among competing goods—choices that have always been present but too little acknowledged. In the end, choosing everything amounts to choosing nothing.

REFERENCES

1. A. Ostebee and P. Zorn, *Calculus from Graphical, Symbolic, and Numerical Points of View*, Volumes 1 and 2, Saunders College Publishing, Philadelphia, 1997.

St. Olaf College
Northfield, MN 55057
ostebee@stolaf.edu
zorn@stolaf.edu

News Items from the Monthly twenty-five years ago ...

Associate Professor G. L. Alexanderson, University of Santa Clara, has been promoted to Professor. p. 818

In spite of this, the Association will not increase dues for 1973. They will remain at $12.50 for individual members. p. 580

... vol. 79, 1972

Rethinking Calculus: Learning and Thinking

James J. Kaput

1. REMODELING CALCULUS THE INSTITUTION. Surely the renewal of Calculus is a good idea, one good enough to attract the attention and energy of many good people. But this is *Calculus the Institution*—that peculiarly American academic event and all its supporting structures and expectations. Professor Knisely, however, barely hints at matters of institutional implementation, so I conclude that he is addressing *Calculus, the System of Knowledge and Technique*. As such, his paper is, perhaps, a warm-up exercise to a deep and long overdue reconsideration of the appropriate intellectual content of Calculus, one that has been postponed while we attempt to remodel Calculus the Institution.

This remodeling has proven to be an arduous task for two reasons: (1) the renovation is taking place whilst the owners and stakeholders continue to inhabit the institution (a constraint applying to most educational reform); and relatedly (2) we have left all the larger structural features of the institution intact, including those features that connect it to the outside world, e.g., to the rapidly changing K-12 education. *The basic architecture and its place in the larger world are untouched.* I suggest that we embark on the more fundamental rebuilding towards which Knisely points. In so doing we need to come to terms with the *relations*, existing and possible, between Calculus the Institution (C-INST) and Calculus the System of Knowledge and Technique (C-KNOWL). And we need to look more deeply and critically at the assumptions, largely tacit, that hold the status quo in place and provide some concrete, implementable alternatives.

2. RELATIONS BETWEEN C-KNOWL AND C-INST. The key relation of interest to me involves learning and cognition. How can ideas and techniques become knowable and usable by those who need to know and use them? And who needs to know them, and in what ways do they need to know them? But before we can get to these questions, we must review the other key relation between C-KNOWL and C-INST, the historical one.

C-INST is the product of several centuries of evolution. The curriculum and texts are rooted in C-KNOWL, which developed at the hands of masters in the 17th and 18th centuries. Many basic curricular structures set down in textbooks by L'Hopital, the Bernoulli's, Euler, and their contemporaries, have remained largely invariant through the 20th century—for a very good reason: they served traditional purposes and populations extremely well. Indeed, this presentation of C-KNOWL is at the foundation of our civilization's scientific and technological infrastructure. While C-KNOWL evolved into an almost sacred academic tradition [**7**], the ambient societies, the nature of education, and the relations between education and the larger society, including and especially in the United States, changed and continue to change profoundly. It is worth noting that, according to Department of Education statistics [**13**], the percentage of students taking AP Calculus today is equal to the percentage graduating from high school a century ago! And, as recently as the 1950's, immediately prior to the huge increases in US access to higher education, calculus was commonly preceded by a preparatory course even at

elite universities. Our expectations regarding who can learn what surely change with the times. The C-INST we know today, while connected to a venerable C-KNOWL, is a relatively recent artifact. Its increasing dysfunction and ill-fit with the new circumstances, especially technological ones, are what gave rise to the Calculus reform movement—the remodeling of C-INST.

3. CALCULUS REFORM FOR THE OTHER 90%. The fixing of C-INST serves only 10% or so of our population, the socio-economic and intellectual elite. The population at large—4 million in each age-level cohort—continues to be denied access to the key ideas of C-KNOWL, a quietly accepted national oversight, despite the fact that we collectively spend billions of dollars supporting a curriculum largely aimed at calculus [**9**]. At the same time we congratulate ourselves on increasing percentages of students passing AP Calculus—to perhaps 3–4% of a given grade-level cohort [**13**]. Not only do we ignore 90% of the population's calculus learning, the 10% on whom we do focus actually need to learn much more mathematics of change and variation than C-INST currently allows them to encounter, for example, the mathematics of dynamical systems: the ideas, representations, and skills needed to make sense of nonlinear phenomena—the mathematics that flourishes in the computational medium [**1**], [**11**]. And perhaps even some of the ideas Knisely suggests.

This is the background against which I wish to discuss learning and thinking of the mathematics of change and variation, including calculus. It is not enough to toss around names of ideas, procedures, and relationships that exist in the formal cultural record of mathematical achievement and in some form or other in the minds of that super-elite constituting professional mathematicians. As the late Morris Kline reminded us, the post-hoc logical structures of these mathematical products may have little to do with the structure of the experiences that students need in order to build viable versions of them in their minds [**5**]. The kind of idle speculation or assertion about the former, without regard for the latter, that appears in Knisely's article represents a general and entirely understandable tendency in our community to think and plan in terms of the cultural artifacts and language that we inhabit. The alert reader will have noticed that I indulged in a bit of this in the previous paragraph when mentioning dynamical systems. As epistemological Flatlanders, we don't distinguish between "up" and "North." But at some point, preferably earlier rather than later, our analyses must turn to learning and thinking, to the conceptual, cultural, and experiential roots of our mathematics—we must break our mind-forged manacles and look *up*.

Each word or phrase that we use to denote some mathematics is but a pointer to a thick web of ideas and relationships, treasures hard-won by great mathematical minds and requiring even greater struggles to be understood by more ordinary minds. I suggest that when we take learning and thinking seriously, we quickly dig to the foundations of our discipline. Simultaneously, we begin the rethinking that reaches beyond remodeling C-INST to build the extended means by which the neglected and under-served 90% may come to know some of the classical C-KNOWL, by which our favored (and important) 10% may come to know it more deeply, and by which both groups may come to learn an even broader and richer mathematics of change and variation. Such rethinking requires an openness to new organizations of ideas, new notations and ways of acting on notations, new uses of interactive technologies that reach beyond the CAS's designed to facilitate or supplant traditional adult competencies with formal symbols, new time scales for learning big ideas (years instead of months), and an enriched conception of what

counts as legitimate mathematical thinking. Our current work is beginning to shed light on the transformative power of such ideas as change and variation to contextualize and organize many of the ideas and skills in K-12 mathematics already regarded as important, and to reveal the efficiencies in curricular organization that are required in order to make room for the new mathematics needed by students living their lives into the second half of the 21st century.

4. RETHINKING CALCULUS—AN ILLUSTRATION. To illustrate I sketch briefly some approaches developed in the ongoing SimCalc Project, with no pretense that it is complete or definitive. We began with a combination of historical analysis that examined attempts by the Scholastics to mathematize change before algebra was available [**4**], the large literature on students' difficulties with kinematics [**8**] and graphs [**6**], and a view that new technologies could be used to support learning that is more foundational than learning facility with traditional notations. Rather, we wished to build the ideas to which these notations conceptually refer, the ideas that they are "about." We asked how the key underlying ideas of rate of change, accumulation, the connections between variable rates and accumulation, and approximation all might be made sensible to as young and diverse a population as possible. This initially meant the middle school grades, but has more recently involved the elementary grades. Following the historical lead and recognizing that the language and metaphors of motion are used quite generally to describe change and variation [**2**], we focused (although not exclusively) on mathematizing linear motion, particularly by controlling motion simulations in familiar or fanciful situations: elevators, people walking or dancing, cars, duckies on a pond, boats in a river, space-vehicles, and so on [**10**].

Our starting criteria were to begin with students' intuitive experience with velocity, to minimize computational complexity, and yet to maintain sufficient variation to avoid the conceptual degeneracy of constant velocity [**12**]. These criteria led to extensive use of piecewise constant velocity functions. Furthermore, we wanted to support direct graphical manipulation of these velocity functions—after all, defining and manipulating piecewise constant functions algebraically is a very cumbersome process.

These considerations lie behind the situation depicted in Figure A, where the graphs on the right side of the figure drive what we usually call "jerky elevators" on the left. Here the student is dragging vertically at the arrow a constant velocity segment (currently at height one floor/sec) in order to make the elevator named "Right" get to the same floor at the same time as elevator "Left," which has a discontinuously decreasing velocity (step) function. In this particular case, if "snap-to-grid" were turned on, forcing all values of time and velocity to be integers, the student could not succeed (9 floors in 7 seconds). Further, the Mean Value Theorem's continuity hypothesis is violated, of course, and its conclusion fails. If a linear velocity function had been used instead of the staircase, then the student would be building an instantiation of Merton's Theorem [**3**, p. 86].

Although it is difficult to build a case on such a narrow shard of curricular activity, this mathematically mundane problem-situation illustrates several aspects of rethinking subject matter as it is experienced and learned by students. First of all, three key underlying ideas—constant rate, mean-value, and area under a rate-graph—are directly and enactively addressed at a level sensible for upper elementary age students. Second, these ideas are approached graphically rather than algebraically, with a tight referential relationship to motion phenomena. Velocity, position, and acceleration graphs in this approach are not only linkable to

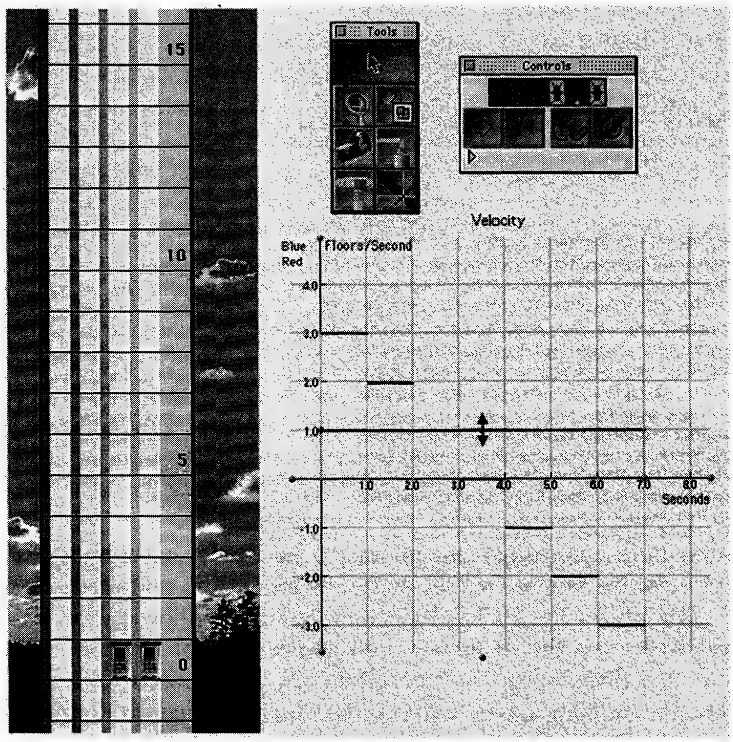


Figure A

each other (or to tables or equations), they also provide three different descriptions of readily viewable and controllable phenomena. That is to say, unlike much school activity based on the "Big Three" representations (numeric, graphic, and algebraic), *they represent something other than each other*! Their primary referential relation is to the phenomena. Third, accumulation, via simple arithmetic sums, is addressed before the subtle ideas of rate and slope.

We could also approximate the staircase by dragging a linear function into place —another interesting reversal of the usual direction, which is to approximate continuously varying functions by discretely varying ones. In our case the approximation-error can be directly computed and predicted by 5th graders, and can be tested by observing the final (or intermediate) positions of the elevators.

5. AN EARLY START: BUILDING ON KINESTHETIC EXPERIENCE. Young children (grades 2–5) first meet mean values in physical activities. Students have pre-quantified notions of their own "slow," "medium," and "fast" speeds that they enact on a marked section of classroom floor. A pair of students is to move along the marked line such that one (A) moves for 2 seconds at "slow" speed, then 2 seconds at "fast," and then 2 seconds again at "slow;" the other student (B) is to move at a constant "medium" such that she reaches the endpoint at exactly the same time as A. Perhaps after a practice run where B tries, at roughly constant speed, to cover the distance that A traveled in 6 seconds, they try to perform this task in parallel. Great struggles ensue as B tries not to be influenced by the fact that at first she is ahead of A and then is behind A. The observing students shout "Constant speed!" "No changing!" and so on, while B fights her own kinesthetic

sense to slow down, to catch up, etc. She is learning, at a very fundamental physical and personal level, a version of constant rate, and perhaps less directly, a sense of average speed, that serves as a foundation for understanding constant functions and linearly increasing distance in their various more formal representations to come, including those encountered in simulations such as in Figure A. Indeed, children's physical experience is difficult for them to quantify but is kinesthetically rich, whereas simulations can be made quantitatively rich, although are kinesthetically vacuous. Hence we engage students in parallel activities on the computer, where they begin with piecewise constant velocity functions labeled only as "slow," "medium," and "fast" (one, two, and four floors/sec, respectively), and gradually move to more quantitative problems and methods, first graphical and numeric and, eventually, algebraic.

Another connection with physical motion is available through the use of motion sensors, which in microcomputer based labs have been used to import and then graph quantitative aspects of phenomena, approximate the data via curve-fitting techniques, etc. In SimCalc MathWorlds **[10]**, it is possible to attach the motion-data to an object and replay it, or edit the motion, or, as was done in Figure B, create a series of motions that relate in some interesting way to the original motion. Here the Frog-character, with the dashed position graph representing a student's actual imported motion, is leading a "Clown Parade" where the clowns were given position functions synthetically. Note the "class-clown" outlier who is marching to her own drummer. (Software graphics are in color, allowing color-coded graphs and actors). The next generation of this software will support uploading of functions from hand-held devices, so each student in the class can control a character in the simulation—they can *be in* the parade!

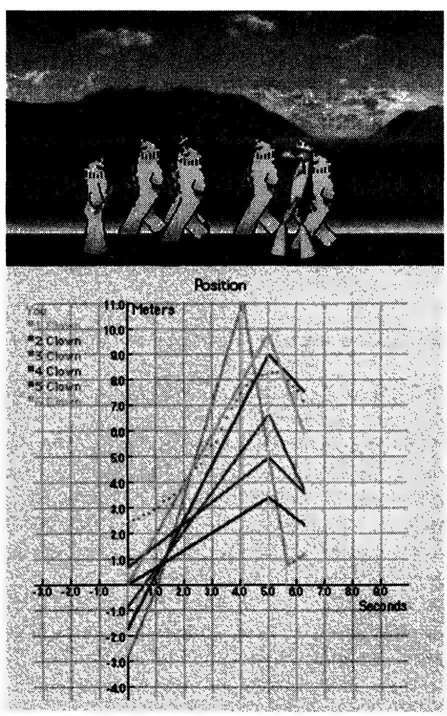


Figure B

6. ONGOING INVESTIGATIONS: FORMALIZATION. An important question is how to get from the informal understandings discussed in preceding sections to a more formal calculus that supports the symbolic technique that *some* students need. We closely study student learning and thinking over extended periods as the students solve problems and work in computer environments of the types described earlier. Their work both lays the base for, and then actually builds, the many ideas at the heart of C-KNOW. As such ideas are being solidly established, their formalization, which is the source of enormous power for those who have mastered it but a great difficulty for many students at all levels, becomes a much more tractable matter. Two formalization strategies are available. One involves beginning with an extended, graphical pre-algebraic mathematics of change experience in the earlier grades and then superimposing algebraic notation upon what is already understood graphically, representing the graphs and phenomena via the usual classes of functions. The second involves co-learning algebra and the ideas that it can represent and manipulate, including rule-based descriptions of motion and change. The second strategy, which perforce must begin in the early grades, introduces and reveals the computational power of the formalisms early and often —as was the case historically. Both strategies provide an experiential anchor for some of the basic functions, their derivatives and integrals. And both offer a major departure from remodeling C-INST for the privileged 10% to building new curricular structures to include ALL students.

The combination of factors that led to Calculus reform applies across the mathematics curriculum, as does the inadequacy of top-down, university-centric reforms. Just as previously successful rote-based learning and coping strategies prove inadequate for students as the mathematical challenges become more substantial, inherited curricular strategies and technologies prove inadequate in the face of our historic challenge to teach much more mathematics to many more people than ever before. And, to go a step farther, our reform strategies may prove inadequate as well. We now must reach beyond the confines of the university, and draw upon deep and detailed understandings of mathematical learning and thinking across many age levels and types of students. This is neither as easy nor as convenient as speculation about neat new topics, but without more foundational work our, and our students', real alternatives will remain limited.

ACKNOWLEDGMENT. SimCalc Project work reported in this paper was supported by NSF Grant # RED 9353507, but views expressed are those of the author and not necessarily those of the NSF.

REFERENCES

1. J. Casti, *Would-be worlds: How simulation is changing the frontiers of science*, John Wiley & Sons, New York, 1996.
2. M. Claggett, *Nicole Oresme and the medieval geometry of qualities and motions, Chapters I & II*, University of Wisconsin Press, Madison, WI, 1968.
3. C. Edwards, *The historical development of the calculus*, Springer-Verlag, Inc., New York, NY, 1979.
4. J. Kaput, Democratizing access to calculus: New routes using old roots, in: *Mathematical thinking and problem solving*, A. Schoenfeld, Editor, Erlbaum, Hillsdale, NJ, 1994, pp. 77–156.
5. M. Kline, Logic vs. pedagogy, *Amer. Math. Monthly* 77 (1970), 264–282.
6. G. Leinhardt, O. Zaslavsky and M. Stein, Functions, graphs, and graphing: Tasks, learning, and teaching, *Review of Educational Research* 60 (1990), 1–64.
7. S. MacLane, Calculus is a discipline, *College Math. J.* 15 (1984), 375.
8. L. McDermott M. Rosenquist and E. v. Zee, Student difficulties in connecting graphs and physics: Examples from kinematics, *Amer. J. Physics* 55 (1987), 503–513.
9. A. Ralston, Computer science, mathematics and the undergraduate curricula in both, *Amer. Math Monthly* 88 (1981), 472–485.
10. SimCalc MathWorlds, software and associated materials available at http://www.simcalc.

umassd.edu.
11. I. Stewart, Change, in: *On the Shoulders of Giants: New approaches to numeracy*, L. Steen, Editor, National Academy Press, Washington, DC, 1990, pp. 183–219.
12. W. Stroup, Embodying a nominalist constructivism, Unpublished doctoral dissertation, Harvard Graduate School of Education, 1996.
13. U.S. Department of Education, *The Condition of Education 1996, Supplemental Table 29-6*, (On-line) National Center For Education Statistics, Washington, DC, 1996. Available at http://www.ed.gov/NCES/pubs/ce/c9629d06.html.

Department of Mathematics
University of Massachusetts–Dartmouth
No. Dartmouth, MA 02747-2300
jkaput@umassd.edu

Problems from the MONTHLY fifty years ago...

4235. *Proposed by Irving Kaplansky, University of Chicago, and D. C. Lewis, University of New Hampshire*

Show that the determinant

$$\begin{vmatrix} (x-1)/1 & (x^2-1)/2 & \cdots & (x^n-1)/n \\ (x^2-1)/2 & (x^3-1)/3 & \cdots & (x^{n+1}-1)/(n+1) \\ \cdots & \cdots & \cdots & \cdots \\ (x^n-1)/n & (x^{n+1}-1)/(n+1) & \cdots & (x^{2n-1}-1)/(2n-1) \end{vmatrix}$$

is a constant times $(x-1)^{n^2}$.

4176 [1945, 522]. *Proposed by H. S. M. Coxeter, University of Toronto*

Prove the following two theorems in affine geometry of three dimensions:

(a) If all the faces of a convex polyhedron are parallelograms, their number is the product of two consecutive integers.

(b) If each face of a convex polyhedron has a center of symmetry, the whole polyhedron has a center of symmetry.

E735 [1946, 394]. *Proposed by Paul Erdős, Stanford University*

Six points can be arranged in the plane so that all triangles formed by triples of these points are isosceles. Show that seven points in the plane cannot be so arranged. What is the least number of points in space which cannot be so arranged?

4252. *Proposed by Paul Erdős, Syracuse University*

It is well known that $2n!/n!(n+1)!$ is always an integer. Prove that for every k there are infinitely many n's such that $2n!/n!(n+k)!$ is an integer.

E768. *Proposed by Irving Kaplansky, University of Chicago*

A number n has the property that for any $p < q < n$,

$$S = p + (p+1) + \cdots + q$$

is never divisible by n. Show that this is true if and only if n is a power of 2.

...vol. 54, 1947

What Do We Do About Calculus? First, Do No Harm

Richard Askey

In memory of Chih-Han Sah

In the spring of 1994, the Dean of our Engineering School paid the expenses of four speakers to tell the Mathematics Department how to teach calculus in a modern way. To him, a modern way was intensive use of computers. The real goal was to have us teach the same amount of calculus but with fewer credits, so that more of the time of engineering students could be spent taking engineering courses. A joint committee was set up to look at what has been taught and what needed to be taught. The conclusion was that with heavy use of computers it would take more time to teach the same material rather than less, so nothing came of the push to cut the number of credits.

I asked the first speaker about proofs. He replied that calculus was not the place to do proofs. Proofs should start in the junior year of college, primarily for students who are mathematics majors.

There was an education meeting held at the University of Chicago. A Japanese education official said, "About half of the ninth-graders could express quantitative relations using letters (variables) and could write geometrical proofs" [**8**]. I asked if ninth grade Japanese students could learn how to do proofs, why couldn't our calculus students also do this? While we have a higher percentage of students taking calculus now than we did forty five years ago when I was a young college student, we do not begin to have 50% of the age cohort taking calculus. Saying that we have so many more students taking calculus that we cannot possibly expect them to be able to do this is looking at the wrong comparison group. In addition to proofs in geometry, there are other proofs in Japanese middle school books, such as a proof that the square root of 2 is irrational. See [**7**].

The second speaker talked about differential equations, and began with this equation:

$$x' = x^2 - t$$

with an initial condition. Once this was put up on an overhead, I worked out the solution. The speaker said this was a differential equation that could not be solved [exactly], and I let this go by without saying anything. After talking on other topics, the speaker came back to this, her favorite equation, and put up an overhead, which seemed to show a pole. She said that a pole was there and that a colleague had shown this. This equation is just a Riccati equation, so can be linearized, and the linear equation solved. In the present case, the linear equation is the Airy equation, so a solution is easy to find. Riccati equations are important in control theory, and there were electrical engineers in the room, so I did not want them to think that the mathematicians did not know what was happening. I asked if they then solved the equation to explain where the pole comes from. She repeated what she had said earlier about a colleague having proven that the equation could not be solved. I said that it depends on what is meant by solved. In the present case, the solution is easily found in terms of Airy functions, which are Bessel functions

in a slight disguise. The pole that was claimed to exist comes from the smallest zero in the denominator.

There is a tendency to downplay the role of infinite series in calculus and in differential equations. The usual argument for differential equations is that it is hard or impossible to see the long range behavior from a power series. In the present case, it is the smallest zero that is in question. The other argument given against power series solutions of differential equations is that few differential equations have solutions that can be written in the form of a nice series. It is one of those miracles of nature that many very important problems lead to just those differential equations that can be solved in series with nice coefficients. These series have the property that the term ratio of the coefficients is a rational function of n, and are called hypergeometric series. A course in calculus is not the place to study hypergeometric series in detail, but the most important one, the binomial theorem, should be there. The ratio test for convergence is as popular with students as it is because it is easy to compute the limit of a term ratio that is a rational function of n, and many of the elementary functions studied in calculus have power series of this type. Students should start to be led in the direction of seeing that this class of functions is important.

The third speaker was someone I have known for years, so I asked some questions in an e-mail before he arrived. One was about differentiating x^n. This can be done in several different ways. The traditional one in our texts was to quote the binomial theorem to get started. This used to be a standard topic in algebra. One of the new calculus books does it this way, and refers the reader to any high school algebra book for a proof of the finite binomial theorem. I called and asked one of the authors if he had looked at any high school algebra or precalculus books recently. He said no. He should, for the binomial theorem is no longer the staple it once was. For example, the precalculus book written by the faculty of The North Carolina School of Science and Mathematics [**1**] does not have either the binomial theorem or the geometric series. In response to my question to readers of the e-mail discussion group calc-reform, someone replied that most of his students had taken calculus in high school. If there is anything students remember from high school, it is the formula for the derivative of x^n, so he does not give a derivation. Many people who are supporting the current reform efforts do not like formulas, and so do not want to use the binomial theorem to differentiate x^n. One solution to this problem is to use another formula. Instead of writing

$$\frac{(x+h)^n - x^n}{h}$$

it is possible to write

$$\frac{y^n - x^n}{y - x}$$

or even

$$\frac{(qx)^n - x^n}{qx - x} = x^{n-1}\frac{q^n - 1}{q - 1}$$

and use the sum of a finite geometric series. This argument also works when n is rational. In the course of changing variables to see this, you give an introduction to the chain rule and to the simple form of l'Hospital's rule.

However, some of those who do not like formulas even object to the formula for the sum of a geometric series, so there is a way to differentiate x^n without using any formula. Just observe that

$$(x+h)^n = (x+h)(x+h)\cdots(x+h)$$

and observe that x^n appears once. The next term, hx^{n-1}, appears once for each factor, so n times. Every other term in the expansion has at least two factors of h. This way the student can understand why

$$(x+h)^n = x^n + nhx^{n-1} + \text{terms that involve } h^2 \text{ or higher powers of } h.$$

Contrast this with the treatment in [**6**]. The formula for the expansion of $(x+h)^n$ is stated when $n = 2, 3, 4, 5$. Then the authors write "we can say that $(x+h)^n = x^n + nx^{n-1}h$ + terms involving h^2 and higher powers of h". There is a big difference between "we can say that" and "we see why". Mathematics should be an open subject, where students do not take such simple facts because "we can say that" or because the computer algebra system gives such a formula.

Another of my concerns can be illustrated by a problem in the same book, but other books could have been used equally well. This deals with when something has been shown to be true.

Consider problem 48 on page 365 of [**6**]. This has three parts. The first is to use Riemann sums to evaluate the integral from 1 to 2 of $\ln x$. The second is to evaluate this integral using anti-derivatives. The same integral had been done in the text, but from 2 to 3. Both of these parts are fine, except it would have been better not to use "evaluate" in the first part, but "approximate", and it would have been better for the students to have been asked to do an integral that had not been done in the text in the second part. However, it is the third part that bothers me. The students are asked to "Explain in words why your answers verify the Fundamental Theorem of Calculus". This has not "verified" the Fundamental Theorem of Calculus, but has illustrated that the approximation in the first part gives an approximation to the exact value obtained in the second. The first definition of "verify" in the dictionary at my desk is: "To prove to be true". Words mean things and they are important. Meanings should not be changed without very good reasons.

The fourth speaker tried to show us how a computer algebra system could be used in a lecture setting. One of his main examples was Simpson's rule. He set up the problem, got three linear equations in three variables, and said that these were far too complicated to solve at the board. He then displayed the solution via a computer algebra system. I asked him why he felt it was necessary to do the interpolation at the points a, $a+h$, and $a+2h$, when you could do the interpolation at $-h$, 0, and h, or even -1, 0, and 1. Then the equations fall apart and it is very easy to do the algebra by hand. There is an important mathematical lesson taught when doing this: you can adapt the coordinate system to the problem at hand. That lesson needs to be learned whether you do calculations by hand or by computer algebra.

In talks on mathematics education, I frequently start with four guidelines that should be considered when teaching, writing a book, or developing a curriculum:

- Do not lie to your students but don't tell them the full truth.
- Some results in mathematics are more important than others and this should be reflected in texts and in class.

- Mathematics is not a secret guild where something is true because I say it is or because a computer algebra system says it is. When something simple and important is studied, reasons should be given.
- Words are important and their meanings should not be changed without very good reasons.

Examples where these have not been followed have already been given. There are many more in newer books. Since first mentioning these, I have decided to add one, which is very important for textbook writers and curriculum developers to observe.

- Be careful that what you are doing does not lead others to make changes that will hurt the long-term education of students.

In his book [**5**] on textbooks written for the TIMSS study, Geoffrey Howson makes the following point: "The passing of the 1960s emphasis on algebraic structure need not be regretted. What is sad is that it has not been replaced by some other clear philosophical or pedagogical structure more appropriate to school mathematics." He ends this paragraph with "A first attempt to establish such a framework of 'recurring themes' has been made by Gardiner [**4**]. It is an idea which deserves further consideration, development, and elaboration."

In the absence of such guidelines, textbook writers, curriculum developers and test writers will look at the current curriculum developers and test writers will look at the current curriculum and try to provide material that will get students ready for later courses. Thus, one frequently overlooked point is how changes being made for one reason will impact in other ways. The newer calculus books tend to be more qualitative, and this is starting to show up on the AP Calculus exams.

For various reasons, which will not be listed here, the knowledge of arithmetic and algebra that students starting calculus have has fallen. As a response to this poorer knowledge of algebra, the Harvard Consortium has tried to finesse the problem by emphasizing the use of graphing calculators. Other reasons are given for this, but a quotation from Tony Phillips suggests that this was a major factor. After saying that students' manipulative skills have become much weaker, Phillips continued with: "And the HCC curriculum makes a great virtue out of this necessity. By eliminating some of the symbolic manipulation from calculus, they were able to make the course more accessible to students." This was written in a newsletter from the Harvard Calculus Consortium.

The report from a committee looking at the future of the AP Calculus exam reads like a description of the Harvard Calculus book. This is a very poor idea. Let me explain why with an analogy. When my son was in high school, the precalculus book they used was [**3**]. This was a very nice book, and he learned a lot from it. We could not use it for the corresponding course at the University of Wisconsin. Our course went about twice as fast, and the students who took it were in general not as good at mathematics as those who took this in high school. Many of the students who will use mathematics seriously are now taking calculus in high school. They need to develop technical skills beyond those of students who take calculus for the qualitative ideas there. The past AP Calculus exams were reasonable exams. My finals in first and second semester calculus were in general a bit harder than the AB and BC exams. A few years ago, there were a couple of students in calculus lectures who wanted to transfer to California Institute of Technology. Exams were given to these students, in chemistry, mathematics, and physics. The mathematics exams over a three year period were sent to me, to share with the students or for

them to take. They were a bit harder than the final exam in my second semester calculus course, as is appropriate. They should be at the level of exams for an honor section. The MAA has published translations of university entrance exams for some Japanese universities [**11**]. The one for students who want to study humanities at Tokyo University is harder than the sophomore placement exam for Cal. Tech. Thus, the level of the traditional AB and BC advanced placement exams was not too high. It is likely that the newer one will continue at an equally high level for a couple of years, but from then on it needs to be watched carefully. There are experienced high school teachers who feel the earlier exams were harder than those given in the last few years, and with a more qualitative exam it will be very easy to have the level slip as students who take the exam have less technical skill.

The message from the Calculus Reform programs that is being heard is that students do not need to be able to do algebra well. The message from the NCTM Reform is that students do not need to know how to do arithmetical calculations well. Both of these messages are different from those sent by the countries that did best on TIMSS. See [**2**] and [**9**].

Technology has a place in mathematics instruction, but it needs to be used carefully. Until much more is learned about the drawbacks, it should not be pushed too much. In England, the results on the age 11 exams in the summer of 1995 were so poor that six months later a decision was made that calculators were no longer to be used in one of the two maths exams. A teacher in Milwaukee talked at the Wisconsin Mathematics Council meeting in May 1995. He said that he was probably the first one to use calculators in high school in the Milwaukee area, and also the first to use graphing calculators. However, he now has some serious doubts about the wisdom of using them as much as he had. First, students do not know enough yet to profit from their use to the same extent that teachers could, with their greater knowledge. Second, he looked at books from 30 years before and found that many topics now in second year algebra were in first year algebra then, topics from second year algebra then are frequently done in precalculus now, and quite a few things that were once done in precalculus are not done in high school now. He did not mentioned specific things, but conic sections comes to mind as something that is frequently not done now, either in precalculus or in calculus. The binomial theorem is another. Both belong somewhere, and high school seems the right place. Unfortunately, NCTM put conic sections down for decreased emphasis. Other countries do them and many other things that we do not do.

We need to look seriously at what is being done in the rest of the world, to see what our students could learn if they had a good mathematics program. Then we need to develop one. The most important part of this is not calculus, but elementary school. However, what we do in calculus has an impact on the rest of our mathematics program, so we need to be very careful about what we do and how we talk about what we are doing. The medical advice of "First, do no harm" is good advice for us as well.

REFERENCES

1. B. B. Barrett et al., *Contemporary Precalculus Through Applications: Functions, data analysis and matrices*, Janson, Dedham, MA, 1992.

2. Albert Beaton et al., *Mathematics Achievement in the Middle School Years*, TIMSS International Study Center, Boston College, Chestnut Hill, 1996.

3. Robert Fisher and Allen Ziebur, *Integrated Algebra and Trigonometry With Analytic Geometry*, Second Edition, Prentice-Hall, Englewood Cliffs, N.J., 1967.
4. Anthony Gardiner, *Recurring Themes in School Mathematics*, Birmingham Univ., UK Mathematics Foundation, Birmingham, 1992.
5. Geoffrey Howson, *Mathematics Textbooks: A Comparative Study of Grade 8 Texts*, TIMSS Monograph No. 3, Pacific Educational Press, Univ. of British Columbia, Vancouver, 1995.
6. Deborah Hughes-Hallett, Andrew M. Gleason, et al., *Calculus*, Wiley, New York, 1994.
7. Kunihiko Kodaira, editor, *Japanese Grade 9 Mathematics, 1984*, translation published by Univ. of Chicago School Mathematics Project, Chicago, 1992.
8. Tatsuro Miwa, Mathematics in Junior and Senior High School in Japan: Present State and Prospects, in [**10**], pp. 172–190.
9. William Schmidt, Curtis McKnight, and Senta Raizen, *A Splintered Vision: An Investigation of U.S. Science and Mathematics Education*, Kluwer, Dordrecht, Boston, London, 1997.
10. Izaak Wirszup and Robert Streit, Developments in School Mathematics Education Around the World, *Proc. of UCSMP International Conference on Mathematics Education*, Univ. Chicago, March 1985, NCTM, Reston 1987.
11. Ling-Erl Eileen T. Wu, *Japanese University Entrance Examination Papers in Mathematics*, Mathematical Association of America, Washington, D.C., 1993.

Dept. of Mathematics
University of Wisconsin-Madison
480 Lincoln Drive
Madison, WI 53706
askey@math.wisc.edu

Personal items from the MONTHLY fifty years ago...

Dr. Paul Erdős has been appointed to a research professorship at Syracuse University.

Associate Professors Garrett Birkhoff and Saunders MacLane have been promoted to professorships.

Professor Saunders MacLane of Harvard University has been appointed to a professorship at the University of Chicago.

The following have received Guggenheim fellowship appointments: Professor Warren Ambrose of the Massachusetts Institute of Technology; Professor Garrett Birkhoff of Harvard University; Professor P. R. Halmos of the University of Chicago; Professor Saunders MacLane of the University of Chicago; Professor A. H. Taub of the University of Washington.

Associate Professor N. E. Steenrod of the University of Michigan has been appointed to an associate professorship at Princeton University.

Dr. Irving Kaplansky at the University of Chicago has been promoted to an assistant professorship.

Professor Antoni Zygmund of the University of Pennsylvania has been appointed to an associate professorship at the University of Oregon.

Associate Professor Ivan Niven of Purdue University has been appointed to an associate professorship at the University of Oregon.

Associate Professor Magnus R. Hestenes of the University of Chicago has been appointed to a professorship at the University of California at Los Angeles.

...vol. 54, 1947

The Fifty-Seventh William Lowell Putnam Mathematical Competition

Leonard F. Klosinski, Gerald L. Alexanderson and Loren C. Larson

The results of the Fifty-Seventh William Lowell Putnam Mathematical Competition, held December 7, 1996, follow. They have been determined in accordance with the regulations governing the Competition, a contest supported by the William Lowell Putnam Prize Fund for the Promotion of Scholarship, a fund endowed by Mrs. Putnam in memory of her husband. The annual Competition is held under the auspices of the Mathematical Association of America.

The first prize, $7,500, was awarded to the Department of Mathematics at Duke University. The members of the winning team were Andrew O. Dittmer, Robert R. Schneck, and Noam M. Shazeer; each was awarded a prize of $500.

The second prize, $5,000, was awarded to the Department of Mathematics at Princeton University. The members of the winning team were Michael J. Goldberg, Craig R. Helfgott, and Jacob A. Rasmussen; each was awarded a prize of $400.

The third prize $3,000 was awarded to the Department of Mathematics at Harvard University. The members of the winning team were Chung-chieh Shan, Stephen S. Wang, and Hong Zhou; each was awarded a prize of $300.

The fourth prize, $2,000, was awarded to the Department of Mathematics at Washington University, St. Louis. The members of the winning team were Mathew B. Crawford, Daniel K. Schepler, and Jade P. Vinson; each was awarded a prize of $200.

The fifth prize, $1,000, was awarded to the Department of Mathematics at the California Institute of Technology. The members of the winning team were Christopher C. Chang, Hui Jin, and Hanhui Yuan; each was awarded a prize of $100.

The six highest ranking individual contestants, in alphabetical order, were Jeremy L. Bem, Cornell University; Ioana Dumitriu, New York University; Robert D. Kleinberg, Cornell University; Dragos N. Oprea, Harvard University; Daniel K. Schepler, Washington University, St. Louis; and Stephen S. Wang, Harvard University. Each of these has been designated a Putnam Fellow by the Mathematical Association of America and awarded a prize of $1,000 by the Putnam Prize Fund.

The next four highest ranking contestants, in alphabetical order, were Federico Ardila, Massachusetts Institute of Technology; Michael R. Korn, Princeton University; Ovidiu Savin, University of Pittsburgh; and Noam M. Shazeer, Duke University; each was awarded a prize of $500.

The next five highest ranking contestants, in alphabetical order, were Andrew O. Dittmer, Duke University; Michael J. Goldberg, Princeton University; Samuel Grushevsky, Harvard University; Craig R. Helfgott, Princeton University; and Adam W. Meyerson, Massachusetts Institute of Technology; each was awarded a prize of $250.

The next ten highest ranking contestants, in alphabetical order, were Pramod N. Achar, Massachusetts Institute of Technology; Constantin S. Chiscanu, Mas-

sachusetts Institute of Technology; Mike L. Develin, Harvard University; Andrei C. Gnepp, Harvard University; Hui Jin, California Institute of Technology; Carl D. Johnson, Georgia Institute of Technology; Amit Khetan, Massachusetts Institute of Technology; Alexandru A. Popa, Princeton University; Robert R. Schneck, Duke University; and Chung-chieh Shan, Harvard University; each was awarded a prize of $100.

The following teams, named in alphabetical order, received honorable mention: the University of Chicago, with team members Nathan D. Broaddhus, Benjamin M. Cowan, and Christopher D. Jeris; the Massachusetts Institute of Technology, with team members Federico Ardila, Eric Kuo, and Adam W. Meyerson; New York University, with team members Aleksandr Bukharovici, Ioana Dumitriu, and Yevgeniy V. Kovchegov; Queen's University, Ontario, with team members Joanna L. Karczmarek, Michael A. Levi, and Allan J. Roberts; and the University of Waterloo, with team members Jason P. Bell, Kevin Purbhoo, and Soroosh Yazdani.

Honorable mention was achieved by the following twenty-eight individuals named in alphabetical order: Dan E. Angelescu, California Institute of Technology; Jason P. Bell, University of Waterloo; Nathan D. Broaddus, University of Chicago; Christopher C. Chang, California Institute of Technology; Donny C. Cheung, University of Waterloo; Patrick K. Corn, Harvard University; Jacob Eliosoff, McGill University; Galen B. Huntington, Reed College; Christopher D. Jeris, University of Chicago; Daniel B. Johnston, Washington University, St. Louis; Joanna L. Karczmarek, Queen's University, Ontario; Travis J. Kopp, Stanford University; François Labelle, McGill University; Ondřej Lhoták, University of Waterloo; Tamas Nemeth, Macalester College; Jacob A. Rasmussen, Princeton University; Robert Ribciuc, Harvard University; Alex Saltman, Harvard University; Yuliy V. Sannikov, Princeton University; Naoki Sato, University of Toronto; Mark J. Tilford, California Institute of Technology; Jiri J. L. Vanicek, Harvard University; Jade P. Vinson, Washington University, St. Louis; Michael J. Westover, California Institute of Technology; Eric G. Yeh, Harvard University; Jun Zhang, University of Utah; Hong Zhou, Harvard University; and Aleksey Zinger, Massachusetts Institute of Technology.

The other individuals who achieved ranks among the top 98, in alphabetical order of their schools, were: Biola University, Jeffrey J. Hatch; the University of British Columbia, Lawrence P. Tang; the California Institute of Technology, Hanhui Yuan; Case Western Reserve University, Neil A. Rubin; Columbia University, Joseph A. Hundley; Cornell University, Harold O. Fox; Duke University, Johanna L. Miller; the University of Florida, Keith A. Grizzell; Hanover College, Navdeep Jaitly; Harvard University, Davin Chor, Samit Dasgupta, Dmitry L. Sagalovskiy, Scott R. Sheffield, Florin Spinu, Jonathan L. Weinstein; the University of Illinois, Urbana-Champaign, Brad A. Friedman; Lebanon Valley College, Jason C. Lee; the University of Louisville, Marc J. Broering; Macalester College, David B. Castro; Marquette University, Scott P. Kempen; the Massachusetts Institute of Technology, Kelvin L. Cheung, Brian C. Dean, Miroslav Jurisic, Edward D. Lee, Michael R. Tehranchi, Benjamin D. Wieland; the University of New Brunswick, Fai K. Tam; New York University, Aleksandr Bukharovich; Ohio University, David Huebel; Princeton University, Andrew M. Neitzke; Queen's College of the City University of New York, Daniil Khaykis; Rice University, Noah A. Rosenberg, Brian M. Wahlert; the University of Richmond, Ronald A. Walker; Rose-Hulman Institute of Technology, Jamie L. Kawabata; Stanford University, Robert G. Au; the University of Texas, Austin, An T. Nguyen; the University of

Texas, El Paso, Ricardo Alberto Sáenz; Washington University, St. Louis, Mathew B. Crawford, Lawrence P. Roberts; and the University of Waterloo, Richard Hoshino, Derek I. E. Kisman, Alex Y. Lee, Kevin Purbhoo, Ian W. T. Vander-Burgh.

The Elizabeth Lowell Putnam Prize, named for the wife of William Lowell Putnam and "awarded to a woman whose performance on the Competition has been deemed particularly meritorious," is awarded this year to Ioana Dumitriu of New York University. The winner is awarded a prize of $500.

There were 2,407 individual contestants from 408 colleges and universities in Canada and the United States in the competition of December 7, 1996. Teams were entered by 294 institutions. The Questions Committee for the fifty-seventh competition consisted of Mark I. Krusemeyer, Carleton College, chair; Richard K. Guy, University of Calgary; and Michael J. Larsen, University of Pennsylvania; they composed the problems and were most prominent among those suggesting solutions.

PROBLEMS

Problem A-1.

Find the least number A such that for any two squares of combined area 1, a rectangle of area A exists such that the two squares can be packed into that rectangle (without the interiors of the squares overlapping). You may assume that the sides of the squares will be parallel to the sides of the rectangle.

Problem A-2.

Let C_1 and C_2 be circles whose centers are 10 units apart and whose radii are 1 and 3. Find, with proof, the locus of all points M for which there exist points X on C_1 and Y on C_2 such that M is the midpoint of the line segment XY.

Problem A-3.

Suppose that each of twenty students has made a choice of anywhere from zero to six courses from a total of six courses offered. Prove or disprove: There are five students and two courses such that all five have chosen both courses or all five have chosen neither.

Problem A-4.

Let S be a set of ordered triples (a, b, c) of distinct elements of a finite set A. Suppose that:

1. $(a, b, c) \in S$ if and only if $(b, c, a) \in S$,
2. $(a, b, c) \in S$ if and only if $(c, b, a) \notin S$,
3. (a, b, c) and (c, d, a) are both in S if and only if (b, c, d) and (d, a, b) are both in S.

Prove that there exists a one-to-one function $g: A \to \mathbf{R}$ such that $g(a) < g(b) < g(c)$ implies $(a, b, c) \in S$.

Problem A-5.

If p is a prime number greater than 3, and $k = \lfloor 2p/3 \rfloor$, prove that the sum

$$\binom{p}{1} + \binom{p}{2} + \cdots + \binom{p}{k}$$

of binomial coefficients is divisible by p^2.

$$\left(\text{For example, } \binom{7}{1} + \binom{7}{2} + \binom{7}{3} + \binom{7}{4} = 7 + 21 + 35 + 35 = 2 \cdot 7^2.\right)$$

Problem A-6.

Let $c \geq 0$ be a constant. Give a complete description, with proof, of the set of all continuous functions $f: \mathbf{R} \to \mathbf{R}$ such that $f(x) = f(x^2 + c)$ for all $x \in \mathbf{R}$.

Problem B-1.

Define a *selfish* set to be a set which has its own cardinality (number of elements) as an element. Find, with proof, the number of subsets of $\{1, 2, \ldots, n\}$ which are *minimal* selfish sets, that is, selfish sets none of whose proper subsets are selfish.

Problem B-2.

Show that for every positive integer n,

$$\left(\frac{2n-1}{e}\right)^{\frac{2n-1}{2}} < 1 \cdot 3 \cdot 5 \cdots (2n-1) < \left(\frac{2n+1}{e}\right)^{\frac{2n+1}{2}}.$$

Problem B-3.

Given that $\{x_1, x_2, \ldots, x_n\} = \{1, 2, \ldots, n\}$, find, with proof, the largest possible value, as a function of n (with $n \geq 2$), of

$$x_1 x_2 + x_2 x_3 + \cdots + x_{n-1} x_n + x_n x_1.$$

Problem B-4.

For any square matrix A, we can define $\sin A$ by the usual power series:

$$\sin A = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} A^{2n+1}.$$

Prove or disprove: There exists a 2×2 matrix A with real entries such that

$$\sin A = \begin{pmatrix} 1 & 1996 \\ 0 & 1 \end{pmatrix}.$$

Problem B-5.

Given a finite string S of symbols X and O, we write $\Delta(S)$ for the number of X's in S minus the number of O's. For example, $\Delta(XOOXOOX) = -1$. We call a string S *balanced* if every substring T of (consecutive symbols of) S has $-2 \leq \Delta(T) \leq 2$. Thus, $XOOXOOX$ is not balanced, since it contains the substring $OOXOO$. Find, with proof, the number of balanced strings of length n.

Problem B-6.

Let $(a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n)$ be the vertices of a convex polygon which contains the origin in its interior. Prove that there exist positive real numbers x and y such that

$$(a_1, b_1) x^{a_1} y^{b_1} + (a_2, b_2) x^{a_2} y^{b_2} + \cdots + (a_n, b_n) x^{a_n} y^{b_n} = (0, 0).$$

SOLUTIONS. In the 12-tuples $(n_{10}, n_9, n_8, n_7, n_6, n_5, n_4, n_3, n_2, n_1, n_0, n_{-1})$ following each problem number below, n_i for $10 \geq i \geq 0$ is the number of students

among the top 206 contestants achieving i points for the problem and n_{-1} is the number of those not submitting solutions.

A-1 (87, 26, 47, 0, 0, 0, 0, 0, 6, 9, 31, 0)

Answer. We can always accommodate the two squares in a rectangle of area $A = (1 + \sqrt{2})/2$.

Solution 1. Let $x = \cos\theta$, $y = \sin\theta$, $0 \le \theta \le \pi/2$. Then

$$x(x+y) = \cos\theta(\cos\theta + \sin\theta) = \sqrt{2}\cos\theta\left(\frac{1}{\sqrt{2}}\cos\theta + \frac{1}{\sqrt{2}}\sin\theta\right)$$
$$= \sqrt{2}\cos\theta\sin(\pi/4+\theta) = \frac{1}{\sqrt{2}}(\sin(2\theta+\pi/4) + \sin(\pi/4)),$$

which is maximized for $2\theta + \pi/4 = \pi/2$. For this value of θ, $x > y$, so the maximum value we desire is $(1 + \sin(\pi/4)/\sqrt{2} = (1+\sqrt{2})/2$.

Solution 2. Write $X = x$ and $Y = ky$ for some positive constant k yet to be determined. Then

$$x(x+y) = X^2 + \frac{XY}{k} \le X^2 + \frac{X^2+Y^2}{2k}$$
$$= x^2 + \frac{x^2+k^2y^2}{2k} = \frac{(2k+1)x^2 + k^2y^2}{2k}.$$

Now choose positive k so that $2k+1 = k^2$, namely, $k = 1+\sqrt{2}$. Then,

$$x(x+y) \le \frac{k^2(x^2+y^2)}{2k} = \frac{k}{2}.$$

For $x = ky$, $X = Y$, the inequality is actually an equality, and $x > y$. So the maximum value is $k/2 = (1+\sqrt{2})/2$.

Solution 3. Maximizing $x(x+y)$ subject to $x^2+y^2 = 1$ by Lagrange multipliers yields

$$2x + y = \lambda\cdot 2x, \quad x = \lambda\cdot 2y, \quad x^2+y^2 = 1.$$

Multiplying the first equation by y and the second by x eventually yields

$$x + y = x\sqrt{2} \text{ (since } x, y \ge 0).$$

So $y = x(\sqrt{2}-1) < x$, and from $x^2+y^2 = 1$, we get

$$x^2 = \frac{1}{4-2\sqrt{2}},$$
$$x^2(x+y)^2 = \frac{1}{4-2\sqrt{2}}\cdot 2x^2 = \frac{2}{(4-2\sqrt{2})^2},$$
$$x(x+y) = \frac{\sqrt{2}}{4-2\sqrt{2}} = \frac{1+\sqrt{2}}{2}.$$

Solution 4. If A is the maximum value of $x(x+y)$ subject to $x^2+y^2 = 1$, then the hyperbola $x(x+y) = A$ is tangent to the circle $x^2+y^2 = 1$. The asymptotes of this hyperbola are the lines $x = 0$ and $x+y = 0$, so the line $y = (\tan(\pi/8))x$ is

an axis of symmetry of both the circle and the hyperbola and passes through the point of tangency. It follows that

$$y = \sin(\pi/8), \quad x = \cos(\pi/8),$$

and thus,

$$\begin{aligned} x(x+y) &= \cos^2(\pi/8) + \sin(\pi/8)\cos(\pi/8) \\ &= \frac{1}{2}(1 + \cos(\pi/4)) + \frac{1}{2}\sin(\pi/4) = \frac{1+\sqrt{2}}{2}. \end{aligned}$$

A-2 (6, 11, 27, 0, 0, 0, 0, 0, 65, 23, 46, 28)

Solution. Let O_1, O_2 be the centers of C_1, C_2 respectively, and let O be the midpoint of O_1O_2. Then the desired set is the closed annulus (ring) with center O, inner radius 1, and outer radius 2.

To see this, note that if M is the midpoint of a line segment XY as described, then

$$\overrightarrow{OM} = \frac{1}{2}(\overrightarrow{OX} + \overrightarrow{OY}) = \frac{1}{2}(\overrightarrow{OO_1} + \overrightarrow{O_1X} + \overrightarrow{OO_2} + \overrightarrow{O_2Y}) = \frac{1}{2}(\overrightarrow{O_1X} + \overrightarrow{O_2Y})$$

since $\overrightarrow{OO_1} + \overrightarrow{OO_2} = \vec{0}$. Now $\overrightarrow{O_1X}$ can range over all vectors of length 1 and $\overrightarrow{O_2Y}$ can range over all vectors of length 3. Therefore $3 - 1 \le |\overrightarrow{O_1X} + \overrightarrow{O_2Y}| \le 3 + 1$ (by the triangle inequality), and it is easy to see that any vector of length between 2 and 4 can indeed be obtained as the sum of a vector of length 1 and a vector of length 3. (Use the law of cosines if you like.) So $\overrightarrow{OM} = \frac{1}{2}(\overrightarrow{O_1X} + \overrightarrow{O_2Y})$ can be any vector of length between 1 and 2, and the result follows.

A-3 (63, 18, 6, 0, 0, 0, 0, 0, 0, 0, 47, 72)

Solution. The answer is "no"; there need be no such set of five students.

Suppose that each of the 20 students chooses exactly 3 courses (and omits to take exactly 3 courses), but each does it in a different way. This is possible, since $20 = \binom{6}{3}$. Then, for each pair of courses, there will be exactly four students enrolled for both courses (their third choices will all be different—one of the $4 = 6 - 2$ courses other than the pair under consideration). No pair of courses will have five common enrollees. Correspondingly, for each pair of courses, there will be exactly four students enrolled for neither of the two, and no pair of courses will have five common absentees.

A-4 (3, 7, 7, 0, 0, 0, 0, 0, 7, 19, 67, 96)

Solution. Intuitively, one regards A as a subset of a circle and S as the set of triples in counterclockwise order. To obtain a linear order, we have to choose a starting point. Fixing $a_0 \in A$, we define a relation $<$ on A by

(i) For all $b \neq a_0$, $a_0 < b$.
(ii) If a_0, b, and c are all distinct, then $b < c$ if and only if $(a_0, b, c) \in S$.

By (1) and (2), for all $b \neq c$, either $b < c$ or $c < b$, but not both. By (1) and (3), $b < c$ and $c < d$ implies $b < d$. Thus $<$ gives A the structure of an ordered set. Defining $g(a) = |\{b \in A | b < a\}|$, we see that $g(a) < g(b) < g(c)$ implies $a < b$

and $b < c$; if $a = a_0$, $(a, b, c) \in S$ by definition. Otherwise, $(a_0, a, b) \in S$, $(a_0, b, c) \in S$, and the result follows from (1) and (3).

A-5 (4, 4, 2, 0, 0, 0, 0, 0, 0, 13, 40, 143)

Solution. Each binomial coefficient is divisible by p (since p divides the numerator and not the denominator of $p!/(r!(p-r)!)$, so we wish to show that

$$1 + \frac{p-1}{2} + \frac{(p-1)(p-2)}{2\cdot 3} + \cdots + \frac{(p-1)\cdots(p-k+1)}{2\cdot 3\cdots k}$$

is divisible by p. The terms are all integers and we express the sum as a sum of fractions whose numerators are multiples of p and whose denominators are prime to p. The sum is equal to

$$\frac{pc_1}{1!} + \frac{pc_2}{2!} + \cdots + \frac{pc_k}{k!} + \left(1 - \frac{1}{2} + \frac{1}{3} - \cdots + \frac{(-1)^{k-1}}{k}\right),$$

where the c_i are integers and the final parenthesis is, when $p = 6q + 1$, $k = 4q$, equal to

$$\begin{aligned}
&1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{4q} - 2\left(\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{4q}\right)\\
&= \frac{1}{2q+1} + \frac{1}{2q+2} + \cdots + \frac{1}{4q}\\
&= \left(\frac{1}{2q+1} + \frac{1}{4q}\right) + \left(\frac{1}{2q+2} + \frac{1}{4q-1}\right)\\
&\quad + \cdots + \left(\frac{1}{2q+q} + \frac{1}{4q-(q-1)}\right)\\
&= \frac{p}{(2q+1)4q} + \frac{p}{(2q+2)(4q-1)} + \cdots + \frac{p}{3q(3q+1)}
\end{aligned}$$

and, when $p = 6q + 5$, $k = 4q + 3$, equal to

$$\begin{aligned}
&1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{4q+3} - 2\left(\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{4q+2}\right)\\
&= \frac{1}{2q+2} + \frac{1}{2q+3} + \cdots + \frac{1}{4q+3}\\
&= \left(\frac{1}{2q+2} + \frac{1}{4q+3}\right) + \left(\frac{1}{2q+3} + \frac{1}{4q+2}\right)\\
&\quad + \cdots + \left(\frac{1}{2q+q+2} + \frac{1}{4q+(3-q)}\right)\\
&= \frac{p}{(2q+2)(4q+3)} + \frac{p}{(2q+3)(4q+2)} + \cdots + \frac{p}{(3q+2)(3q+3)}.
\end{aligned}$$

A-6 (4, 4, 11, 0, 0, 0, 0, 0, 18, 13, 46, 110)

Solution. We begin with the general observation that $f(x) = f(x^2 + c) = f(-x)$, so f is always even. Conversely, the even extension of any continuous function on

$[0,\infty)$ satisfies the functional equation as long as the original function does. Therefore, we may and do restrict attention to $x \geq 0$ in everything that follows. We consider two cases.

Case 1: $0 \leq c \leq 1/4$.

Here, $x^2 - x + c = 0$ has positive zeros, $a = (1 - \sqrt{1-4c})/2$ and $b = (1 + \sqrt{1-4c})/2$. If $0 \leq x_0 < b$ and we define $x_{n+1} = x_n^2 + c$, the monotonicity of $x^2 + c$ on $[0,\infty)$ implies that $x_0, x_1, \ldots$ is monotonic (increasing for $0 < x_0 < a$, decreasing for $a < x_0 < b$) and bounded, therefore convergent, and therefore convergent to a since the limit must satisfy $L = L^2 + c$. As

$$f(x_0) = f(x_1) = \cdots = \lim_{n\to\infty} f(x_n) = f\Big(\lim_{n\to\infty} x_n\Big) = f(a),$$

we have $f(x) = f(a)$ for all x, $0 \leq x < b$.

If $x_0 > b$, the monotonicity of $\sqrt{x-c}$ guarantees that $x_0 > \sqrt{x_0 - c} > b$, so we can define recursively, $x_{n+1} = \sqrt{x_n - c}$. Again, the sequence (x_n) is bounded and monotonic; therefore it also has a limit, and this limit must be b. Then

$$f(x_0) = f(x_1) = \cdots = \lim_{n\to\infty} f(x_n) = f\Big(\lim_{n\to\infty} x_n\Big) = f(b).$$

As the range of f is finite and f is continuous, it is constant.

Case 2: $c > 1/4$.

Now, $x \to x^2 + c$ has no real fixed points. Setting $t_0 = 0$, $t_{n+1} = t_n^2 + c$, the sequence (t_i) is monotonic, so if it didn't go to infinity, it would have to converge to a (non-existent) fixed point. So each $x \geq 0$ is in some interval $[t_n, t_{n+1}]$.

Let g be any continuous function on the interval $[0,c]$ such that $g(c) = g(0)$. Define $\phi(x) = \sqrt{x-c}$ and

$$f(x) = \begin{cases} g(x) & \text{for } x \in [0,c] = [t_0, t_1] \\ g(\phi(x)) & \text{for } x \in [c, c^2+c] = [t_1, t_2] \\ g(\phi(\phi(x))) & \text{for } x \in [t_2, t_3], \\ \text{and in general} & \\ g\Big(\underbrace{\phi(\phi(\cdots(\phi}_{n}(x))\cdots))\Big) & \text{for } x \in [t_n, t_{n+1}] \end{cases}$$

By construction, $f(x)$ satisfies the desired functional equation. Continuity is obvious except at the points t_i, where it follows from $g(c) = g(0)$. Conversely, every function $f(x)$ is determined by its values on $[0, c)$.

B-1 (113, 72, 17, 0, 0, 0, 0, 0, 3, 1, 0, 0)

Solution. The cardinality is the *least* member of a minimal selfish set, else there would be a proper selfish subset. But there is no other restriction. So their number is

$$\binom{n-1}{0} + \binom{n-2}{1} + \binom{n-3}{2} + \ldots,$$

which, by induction, is the nth Fibonacci number, F_n, where $F_1 = F_2 = 1$, $F_{n+1} = F_n + F_{n-1}$.

B-2 (85, 13, 6, 0, 0, 0, 0, 0, 15, 7, 28, 52)

Solution. Let M be the (natural) logarithm of the product "in the middle":

$$M = \ln(1 \cdot 3 \cdot 5 \cdots (2n-1)) = \ln 3 + \ln 5 + \cdots + \ln(2n-1).$$

If we take twice this quantity, we can interpret that as a Riemann sum for a definite integral of $\ln x$, and therefore we get estimates (using (a) partition points $3, 5, 7, \dots, 2n+1$, together with left-hand endpoints, and (b) partition points $1, 3, 5, \dots, 2n-1$, together with right-hand endpoints):

$$2M < \int_3^{2n+1} \ln x\,dx \qquad \text{and} \qquad 2M > \int_1^{2n-1} \ln x\,dx.$$

So we have

$$\int_1^{2n-1} \ln x\,dx < 2M < \int_3^{2n+1} \ln x\,dx$$

$$[x \ln x - x]_1^{2n-1} < 2M < [x \ln x - x]_3^{2n+1}$$

$$(2n-1)\ln(2n-1) - (2n-1) + 1 < 2M < (2n+1)\ln(2n+1) \\ -(2n+1) - (3\ln 3 - 3),$$

so certainly

$$(2n-1)\ln(2n-1) - (2n-1) < 2M < (2n+1)\ln(2n+1) - (2n+1)$$

$$\frac{2n-1}{2} \ln\left(\frac{2n-1}{e}\right) < M < \frac{2n+1}{2} \ln\left(\frac{2n+1}{e}\right)$$

Taking exponentials now yields the desired result.

B-3 (20, 19, 8, 0, 0, 0, 0, 0, 14, 36, 56, 53)

Solution. Equivalently, we have to form an n-bead string out of beads labelled from 1 to n in order to maximize the sum of products of adjacent bead values. Suppose that for some $k < n$, the set of beads 1 through k form a single connected chain but that bead $k+1$ is not adjacent to any bead in that chain. Then there exists a connected chain of beads that contains all beads 1 through k, but does not contain bead $k+1$, and is such that one end is adjacent to bead $k+1$, while the other is one of the beads from 1 to k. By cutting out this chain and reversing its order, we obtain a new string with greater product sum. Thus, $k+1$ must be adjacent to the string with beads 1 through k. In fact, it must be adjacent to the bead with the smaller value of the two end beads, since otherwise we could increase the product sum by cutting out the string with beads 1 through k and reversing the order. By induction on k, the string reads $\dots, 7, 5, 3, 1, 2, 4, 6, 8, \dots$. We conclude by a standard summation that the answer is $(2n^3 + 3n^2 - 11n + 18)/6$.

B-4 (40, 4, 7, 0, 0, 0, 0, 0, 17, 4, 48, 86)

Solution. We'll show that there is *no* such matrix A. First of all, note that for any invertible matrix P and any square matrix B of the same size,

$$\begin{aligned}\sin(PBP^{-1}) &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}(PBP^{-1})^{2n+1} \\ &= \sum_{=0}^{\infty} \frac{(-1)^n}{(2n+1)!} PB^{2n+1}P^{-1} \\ &= P\left(\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} B^{2n+1} \right) P^{-1} \\ &= P(\sin B)P^{-1}.\end{aligned}$$

Thus the sines of similar matrices are similar.

Now any 2×2 matrix A with real entries is similar to either a diagonal matrix $\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ with real or complex entries λ_1, λ_2, or a triangular matrix $\begin{pmatrix} \lambda & c \\ 0 & \lambda \end{pmatrix}$ with real entries λ, c (in fact, one can take $c = 1$). Therefore, $\sin A$ is similar to either $\sin\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ or $\sin\begin{pmatrix} \lambda & c \\ 0 & \lambda \end{pmatrix}$. But $\sin\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ is a diagonal matrix, since all powers of $\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ are diagonal, and no diagonal matrix is similar to $\begin{pmatrix} 1 & 1996 \\ 0 & 1 \end{pmatrix}$, since the latter is not diagonalizable. So if $\sin A = \begin{pmatrix} 1 & 1996 \\ 0 & 1 \end{pmatrix}$, there must be real numbers λ and c such that $\begin{pmatrix} 1 & 1996 \\ 0 & 1 \end{pmatrix}$ is similar to $\sin\begin{pmatrix} \lambda & c \\ 0 & \lambda \end{pmatrix}$.

Let $U = \begin{pmatrix} \lambda & c \\ 0 & \lambda \end{pmatrix}$. We compute $\sin U$ explicitly: we have $U^2 = \begin{pmatrix} \lambda^2 & 2\lambda c \\ 0 & \lambda^2 \end{pmatrix}$, $U^3 = \begin{pmatrix} \lambda^3 & 3\lambda^2 c \\ 0 & \lambda^3 \end{pmatrix}, \ldots$, and by induction $U^n = \begin{pmatrix} \lambda^n & n\lambda^{n-1}c \\ 0 & \lambda^n \end{pmatrix}$. Therefore,

$$\begin{aligned}\sin U &= \sum \frac{(-1)^n}{(2n+1)!} \begin{pmatrix} \lambda^{2n+1} & (2n+1)\lambda^{2n}c \\ 0 & \lambda^{2n+1} \end{pmatrix} \\ &= \begin{pmatrix} \sum \frac{(-1)^n \lambda^{2n+1}}{(2n+1)!} & \sum \frac{(-1)^n \lambda^{2n}}{(2n)!} c \\ 0 & \sum \frac{(-1)^n \lambda^{2n+1}}{(2n+1)!} \end{pmatrix} = \begin{pmatrix} \sin\lambda & c\cos\lambda \\ 0 & \sin\lambda \end{pmatrix}.\end{aligned}$$

For this matrix to be similar to $\begin{pmatrix} 1 & 1996 \\ 0 & 1 \end{pmatrix}$, the double eigenvalue $\sin\lambda$ must equal 1. But then $\cos\lambda = 0$ and so $\sin U = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, which is not similar to $\begin{pmatrix} 1 & 1996 \\ 0 & 1 \end{pmatrix}$ after all, so we have a contradiction.

B-3 (34, 4, 7, 0, 0, 0, 0, 0, 3, 2, 58, 98)

Solution. Balanced strings consist of X's and O's arranged alternately, or with as many as two consecutive letters of the same kind. Such occurrences of double letters must happen alternately, $\ldots XX \ldots OO \ldots XX \ldots$, with an even number (possibly zero) of single letters between each occurrence. If b_n is the number of

balanced strings of length n, we will show that

$$b_{n+2} = 2b_n + 2, \quad \text{i.e.,}\ b_{n+2} + 2 = 2(b_n + 2).$$

(When we have counted X, O, $b_1 = 2$; XX, XO, OX, OO, $b_2 = 4$; XXO, XOX, XOO, OXX, OXO, OOX, $b_3 = 6$; $XXOX$, $XXOO$, $XOXX$, $XOXO$, $XOOX$, $OXXO$, $OXOX$, $OXOO$, $OOXX$, $OOXO$, $b_4 = 10$; then we have more than enough information to deduce that $b_{2n} = 3 \cdot 2^n - 2$ and $b_{2n-1} = 2^{n+1} - 2$.)

Let x_n, y_n, z_n denote, respectively, the number of balanced strings of length n that end with XX, the number that end with XO and whose last occurrence of a double letter was XX, and the number that end with XO and whose last occurrence of a double letter was OO. Then x_n, y_n, z_n also denote, respectively, the number of balanced strings of length n that end with OO, with OX and whose last occurrence of a double letter was OO, and with OX and whose last occurrence of a double letter was XX. (We count the purely alternating strings $XOX \ldots OX$ and $OXO \ldots XO$ twice, once among the y_n and once among the z_n.) So, $b_n = 2x_n + 2y_n + 2z_n - 2$, or $b_n + 2 = 2(x_n + y_n + z_n)$.

We can form balanced strings of length $n + 2$ from such strings of length n in exactly the following ways:

i. $\ldots XX$; legal to add OO or OX, but nothing else; $(2x_n)$;
ii. $\ldots XO$ with last double letter XX; legal to add OX or XO, but nothing else; $(2y_n)$;
iii. XO with last double letter OO, legal to add XX or XO, but nothing else; $(2z_n)$.

Similarly, with O and X interchanged throughout. Thus, $b_{n+2} + 2 = 4(x_n + y_n + z_n) = 2(b_n + 2)$.

B-6 (0, 1, 9, 0, 0, 0, 0, 0, 2, 0, 23, 171)

Solution. Let $f(\vec{v}) = \sum_i e^{(a_i, b_i)\cdot\vec{v}}$. If $\nabla f(x_0, y_0) = \vec{0}$, then (e^{x_0}, e^{y_0}) is a solution of the original vector equation. As ∇f is a continuous vector field, to prove that it has a zero, it suffices to find a simple closed contour over which ∇f has a nonzero winding number. We claim that any sufficiently large counterclockwise circle around the origin will do. Indeed, for some $r > 0$,

$$\max_i \{\vec{v} \cdot (a_i, b_i)\} \geq r\|\vec{v}\|,$$

so

$$\begin{aligned} \vec{v} \cdot \nabla f(\vec{v}) &= \sum_{i=1}^{n} e^{(a_i, b_i)\cdot\vec{v}}\big((a_i, b_i) \cdot \vec{v}\big) \\ &\geq r\|\vec{v}\|e^{r\|\vec{v}\|} + (n-1)\inf_x xe^x = r\|\vec{v}\|e^{r\|\vec{v}\|} - (n-1)/e. \end{aligned}$$

For $\|\vec{v}\| = R \gg 0$, this means that $\vec{v} \cdot \nabla f(\vec{v}) > 0$, which means that the winding number of $\nabla f(\vec{v})$ along a circular path of radius R is 1.

Klosinski / Alexanderson:
Department of Mathematics
Santa Clara University
Santa Clara, CA 95053-0290

Larson:
Department of Mathematics
St. Olaf College
Northfield, MN 55057

NOTES

Edited by **Jimmie D. Lawson**

A Quadratic Trio

Joseph Kupka

Strictly speaking, the quadratic formula is unnecessary. One may always complete the square. We use and treasure the quadratic formula because it is often less tedious than completing the square.

The purpose of this note is to treat the Maclaurin series of the function

$$f(x) = \frac{\beta x + \gamma}{Ax^2 + Bx + C} = \sum_{n=0}^{\infty} A_n x^n, \quad A, C \neq 0$$

in the same spirit. We refer to f as the *generating function* for the sequence of coefficients A_n. The standard technique for determining the A_n is to decompose f into partial fractions and expand these into (possibly complex) geometric series. We use it, once and for all, to produce three formulas free of complex notation. Illustrations of this "quadratic trio" include calculations connected with the tossing of a coin with probability p for heads and $q = 1 - p$ for tails.

(i) If $B^2 - 4AC > 0$, then

(1)
$$A_n = \frac{(-1)^n}{D}\left[\left(\gamma\left(\frac{B+D}{2C}\right) - \beta\right)\left(\frac{B+D}{2C}\right)^n - \left(\gamma\left(\frac{B-D}{2C}\right) - \beta\right)\left(\frac{B-D}{2C}\right)^n\right],$$

where $D = \sqrt{B^2 - 4AC}$. Note: the $(B \pm D)/2C$ are *not* misprints of $(B \pm D)/2A$. The Fibonacci sequence ($f_0 = 0$, $f_1 = 1$, $f_n = f_{n-1} + f_{n-2}$, $n \geq 2$), which has generating function $F(x) = \sum_{n=0}^{\infty} f_n x^n = x/(1 - x - x^2)$, provides a straightforward illustration of (1). Substituting $A = B = -1$, $C = \beta = 1$, $\gamma = 0$, $D = \sqrt{5}$ into (1) gives

$$f_n = \frac{(-1)^n}{\sqrt{5}}\left[-\left(\frac{-1+\sqrt{5}}{2}\right)^n + \left(\frac{-1-\sqrt{5}}{2}\right)^n\right],$$

which simplifies at once to the more familiar form

$$f_n = \frac{1}{\sqrt{5}}\left[\left(\frac{\sqrt{5}+1}{2}\right)^n + (-1)^{n+1}\left(\frac{\sqrt{5}-1}{2}\right)^n\right].$$

A sometimes simpler alternative to (1) is

$$(1^*) \qquad A_n = \frac{1}{A(r_1 - r_2)}\left[\left(\frac{\gamma}{r_2} + \beta\right)\left(\frac{1}{r_2}\right)^n - \left(\frac{\gamma}{r_1} + \beta\right)\left(\frac{1}{r_1}\right)^n\right],$$

where r_1, r_2 denote the roots of $Ax^2 + Bx + C$. For example, when $p \neq q$, the classical formula $(1 + (q - p)^n)/2$ for the probability of an even number of heads in n tosses of the coin is more easily obtained from the generating function $(1 - qs)/(1 - s)(1 - (q - p)s)$ by using (1*) instead of (1).

(ii) If $B^2 - 4AC = 0$, then

$$(2) \qquad A_n = \frac{(-1)^n}{A}\left(\frac{2A}{B}\right)^{n+1}\left[(n+1)\gamma\left(\frac{2A}{B}\right) - n\beta\right].$$

This is obtained via the identity $(1 - x)^{-2} = \sum_{n=0}^{\infty}(n + 1)x^n$.

(iii) If $B^2 - 4AC < 0$, and if $A > 0$, then

$$(3) \qquad A_n = \frac{2}{D}\left(\frac{A}{C}\right)^{n/2}\left[\gamma\left(\frac{A}{C}\right)^{1/2}\sin(n+1)\theta + \beta\sin n\theta\right],$$

where $D = \sqrt{4AC - B^2}$ and $\theta = \cos^{-1}(-B/2\sqrt{AC})$. When $A < 0$, (3) may be used mutatis mutandis: specifically, multiply the right-hand side of (3) by -1 after changing θ to $\cos^{-1}(B/2\sqrt{AC})$. When $A > 0$, a sometimes more convenient alternative to (3) is

$$(3^*) \qquad A_n = \frac{2R}{D}\left(\frac{A}{C}\right)^{n/2}\sin(n\theta + \psi),$$

where $R = \sqrt{A\gamma^2 - B\beta\gamma + C\beta^2}\,/\,\sqrt{C}$ and $\psi = \operatorname{sgn}(\gamma)\cos^{-1}((\beta - \gamma B/2C)/R)$ under the convention $\operatorname{sgn}(x) = 1$ if $x \geq 0$, and $= -1$ otherwise. For example, when $p = 3/4$, the formula $1 - 2(3/4)^{n+1} + 2(\sqrt{3}/4)^{n+1}\sin((n + 1)\theta - \pi/2)$, $\theta = \cos^{-1}(-1/\sqrt{3})$, for the probability of three heads in a row during n tosses of the coin comes a little more directly from the generating function $p^3s^3/(1 - s)(1 - qs - pqs^2 - p^2qs^3)$ if one uses (3*).

In heavy-duty applications, these formulas can be considerably less tedious than the standard technique. For example, the expression

$$p_n = \sum_{i=0}^{\infty}\binom{n - 2i - 2}{2i}p^{2i+2}q^{n-2i-2}$$

represents the probability that, in n tosses of the coin, the number of heads is even —*and*, that the *first* pair of consecutive heads appears on the *last* two tosses. In this case with $p_0 = p_1 = 0$, we have

$$P(s) = \sum_{n=0}^{\infty} p_n s^n = \frac{p^2s^2(1 - qs)}{(1 - qs)^2 - (pqs^2)^2}.$$

A complete partial fractions decomposition of $P(s)$ would be forbidding, but it is quite easy to express $P(s) = p(P_1(s) + P_2(s))/2q$, where $P_1(s) = (qs - 1)/(pqs^2 + qs - 1)$ and $P_2(s) = (qs - 1)/(pqs^2 - qs + 1)$. The quadratic trio now provides Σ-free formulas for the p_n with relatively little additional algebra to achieve simplest form. Only (1) is required for the case $p < 1/5$, (1) and (2) for $p = 1/5$, and (1) and (3) for $p > 1/5$. It is unexpected to find $p = 1/5$ as a changeover point between substantially different formulas for the p_n. When $p = 1/2$, we have $\theta = \pi/3$, which causes the sequence $\{\gamma\sqrt{A/C}\sin(n + 1)\theta + \beta\sin n\theta\} = (1/2)\{\sin((n - 1)\pi/3)\}$ to "deconstruct" into the period-six sequence

$(\sqrt{3}/4)\{-1, 0, 1, 1, 0, -1, -1, 0, \dots\}$. The overall result is

$$2^n p_n = \sum_{i=0}^{\infty} \binom{n-2i-2}{2i}$$
$$= \frac{1}{2^n\sqrt{5}}\left((\sqrt{5}+1)^{n-1} + (-1)^n(\sqrt{5}-1)^{n-1}\right) + \frac{1}{2}\varepsilon_n,$$

where $\varepsilon_n = -1$ if $n \equiv 0, 5 (\mathrm{mod}\, 6)$; $\varepsilon_n = 0$ if $n \equiv 1, 4 (\mathrm{mod}\, 6)$; and $\varepsilon_n = 1$ if $n \equiv 2, 3 (\mathrm{mod}\, 6)$. It is important to emphasize how *easily* and *quickly* this result is obtained from $P_1(s)$, $P_2(s)$, and the quadratic trio.

In a completely different application of the quadratic trio, we obtain explicit formulas for the nth derivative of

$$f(x) = \frac{1+x}{1+x+x^2} = \frac{(x-a)+(a+1)}{(x-a)^2+(2a+1)x+(a^2+a+1)},$$

where a is any constant. The drill is: Obtain the coefficient A_n of $(x-a)^n$ in the Taylor series of f about a from, in this case, (3), change a into x, set $g(x) = \cos^{-1}(-(2x+1)/2\sqrt{x^2+x+1})$, and multiply by $n!$ to get

$$f^{(n)}(x) = (2/\sqrt{3})n!\,(x^2+x+1)^{-n/2}\left[\frac{(x+1)\sin((n+1)g(x))}{\sqrt{x^2+x+1}} + \sin(ng(x))\right].$$

With (3*) in place of (3), we would get

$$f^{(n)}(x) = (2/\sqrt{3})n!\,(x^2+x+1)^{-(n+1)/2}\sin(ng(x)+h(x)),$$

where $h(x) = \mathrm{sgn}(1+x)\cos^{-1}((1-x)/2\sqrt{x^2+x+1})$.

One may envisage a "cubic quartet" of formulas for the coefficients in the Maclaurin series of the ratio of a quadratic and a cubic. The quadratic trio would facilitate the derivation of such formulas.

Department of Mathematics
Monash University
Clayton, Victoria 3168
Australia

A Discrete Form of the Beckman–Quarles Theorem

Apoloniusz Tyszka

The following theorem may be viewed (for $n = 2$) as a discrete form of the classical Beckman–Quarles theorem, which states that any map from $\mathbb{R}^n$ to $\mathbb{R}^n$ $(2 \le n < \infty)$ preserving unit distances is an isometry; see [**1**].

Theorem. *If $x, y \in \mathbb{R}^2$ and $|x-y|$ is constructible by means of ruler and compass, then there exists a finite set $S_{xy} \subseteq \mathbb{R}^2$ containing x and y such that each map from S_{xy} to $\mathbb{R}^2$ preserving all unit distances preserves the distance between x and y.*

Proof: It is known that a segment can be constructed with the use of a ruler and a compass if and only if its length belongs to the real quadratic closure of the field of

rational numbers [5]. Let us denote by D the set of all non-negative numbers d with the following property:

> If $x, y \in \mathbb{R}^2$ and $|x-y| = d$ then there exists a finite set S_{xy} such that $x, y \in S_{xy}$ and any map $f: S_{xy} \to \mathbb{R}^2$ that preserves unit distance also preserves the distance between x and y.

Obviously $0, 1 \in D$. We first prove that if $d \in D$ then $\sqrt{3} \cdot d \in D$. Let us assume that $d > 0$, $x, y \in \mathbb{R}^2$, $|x-y| = \sqrt{3} \cdot d$. Using the notation of Figure 1, we show that

$$S_{xy} := \bigcup \left\{S_{ab}: a, b \in \left\{x, y, x_1, x_2, \tilde{x}_1, \tilde{x}_2, \tilde{y}\right\}, |a-b| = d\right\}$$

is adequate for the segment xy.

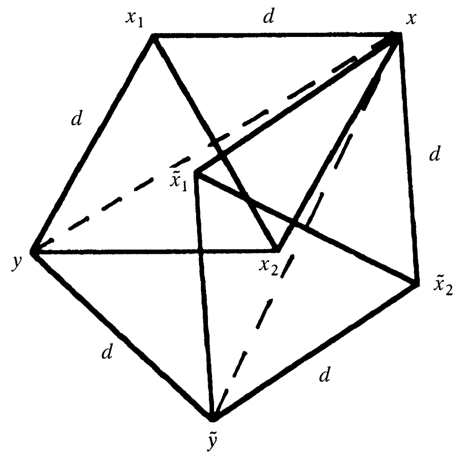


Figure 1

Let us assume that f: $S_{xy} \to \mathbb{R}^2$ preserves the distance 1. Since

$$S_{xy} \supseteq S_{y\tilde{y}} \cup S_{xx_1} \cup S_{xx_2} \cup S_{yx_1} \cup S_{yx_2} \cup S_{x_1x_2},$$

we conclude that f preserves the distances between y and $\tilde{y}$, x and x_1, x and x_2, y and x_1, y and x_2, and x_1 and x_2. Hence $|f(y) - f(\tilde{y})| = d$ and $|f(x) - f(y)|$ is either 0 or $\sqrt{3} \cdot d$. Analogously we have that $|f(x) - f(\tilde{y})|$ is either 0 or $\sqrt{3} \cdot d$. Thus $f(x) \neq f(y)$, so $|f(x) - f(y)| = \sqrt{3} \cdot d$, which completes the proof that $\sqrt{3} \cdot d \in D$.

If $d \in D$, then $2 \cdot d \in D$ (see Figure 2).

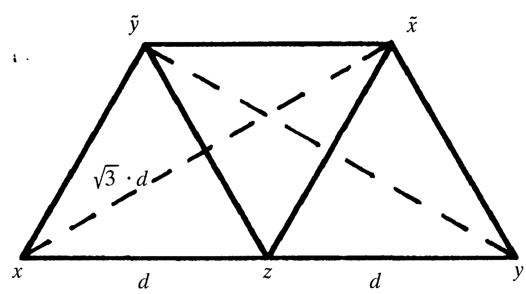


Figure 2.

$$|x-y| = 2 \cdot d$$

$$S_{xy} = \bigcup \left\{S_{ab}: a, b \in \{x, y, \tilde{x}, \tilde{y}, z\},\ \ |a-b| = d \vee |a-b| = \sqrt{3} \cdot d\right\}$$

From Figure 3 it is clear that if $d \in D$ then all distances $k \cdot d$ (k a positive integer) belong to D.

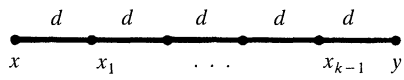


Figure 3.

$$|x - y| = k \cdot d$$

$$S_{xy} = \bigcup \{S_{ab} \colon a, b \in \{x, y, x_1, \dots, x_{k-1}\},\quad |a - b| = d \vee |a - b| = 2 \cdot d\}$$

From Figure 4 it is clear that if $d \in D$, then all distances d/k (k a positive integer) belong to D. Hence $D \supseteq \mathbb{Q}^+$; here and subsequently $\mathbb{Q}^+$ denotes the set of positive rational numbers.

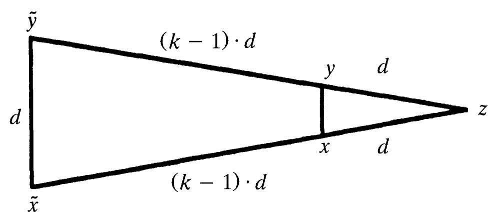


Figure 4.

$$|x - y| = \frac{d}{k}$$

$$S_{xy} = S_{\tilde{x}\tilde{y}} \cup S_{\tilde{x}x} \cup S_{xz} \cup S_{\tilde{x}z} \cup S_{\tilde{y}y} \cup S_{yz} \cup S_{\tilde{y}z}$$

If $a, b \in D$, $a > b$, then $\sqrt{a^2 - b^2} \in D$ (see Figure 5). Hence

$$\sqrt{2} \cdot a = \sqrt{\left(\sqrt{3} \cdot a\right)^2 - a^2} \in D \quad \text{and} \quad \sqrt{a^2 + b^2} = \sqrt{\left(\sqrt{2} \cdot a\right)^2 - \left(\sqrt{a^2 - b^2}\right)^2} \in D.$$

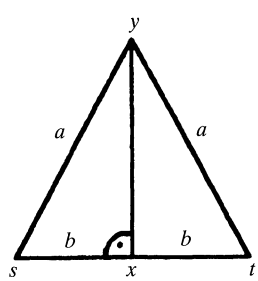


Figure 5.

$$|x - y| = \sqrt{a^2 - b^2}$$

$$S_{xy} = S_{sx} \cup S_{xt} \cup S_{st} \cup S_{sy} \cup S_{ty}$$

The construction presented in Figure 6 shows that if $a, b \in D$, $a > b$, then $a - b \in D$, hence $a + b = 2 \cdot a - (a - b) \in D$.

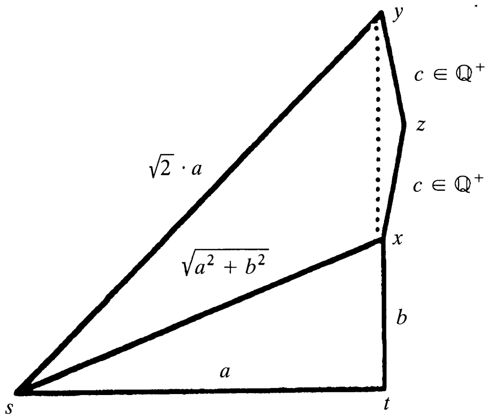


Figure 6.

$|x - y| = a - b, \quad 2 \cdot c < a + b$

$S_{xy} = S_{st} \cup S_{tx} \cup S_{sx} \cup S_{sy} \cup S_{ty} \cup S_{zx} \cup S_{zy}$

In order to prove that $D \setminus \{0\}$ is a multiplicative group it remains to observe that if positive $a, b, c \in D$ then $ab/c \in D$ (see Figure 7).

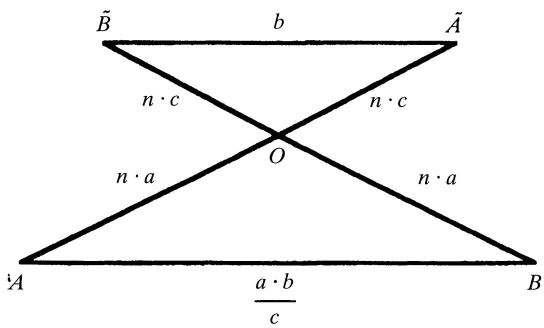


Figure 7.

$b < 2 \cdot n \cdot c$

$S_{AB} = S_{OA} \cup S_{OB} \cup S_{O\tilde{A}} \cup S_{O\tilde{B}} \cup S_{A\tilde{A}} \cup S_{B\tilde{B}} \cup S_{\tilde{A}\tilde{B}}$

If $a \in D$, $a > 1$, then $\sqrt{a} = \frac{1}{2} \cdot \sqrt{(a + 1)^2 - (a - 1)^2} \in D$; if $a \in D$, $0 < a < 1$, then $\sqrt{a} = 1/\sqrt{\frac{1}{a}} \in D$. Thus D contains all non-negative real numbers contained in the quadratic closure of $\mathbb{Q}$.

Remarks. One can easily see that if a distance d belongs to the set D defined in the first paragraph of our proof, then d is definable in $\mathbb{R}$ by some formula in the language of fields; see [**2**, pp. 35, 41] for formal definitions of these terms. From this, we can conclude that any such d must be an algebraic number [**2**, p. 197]. On the other hand, applying results from [**7**] and [**8**] (and from [**9**] for $n > 2$), we can prove that any algebraic distance belongs to D.

In our proof we used some ideas of [**6**], and some ideas are based on constructions of Georg Mohr, who first proved that all Euclidean constructions can be carried out with compass alone; see [**4**] and [**3**]. Using these ideas we recently extended the theorem to constructible differences in $\mathbb{R}^n$ ($2 \leq n < \infty$) and multivalued mappings.

REFERENCES

1. F. S. Beckman and D. A. Quarles Jr., On isometries of euclidean spaces, *Proc. Amer. Math. Soc.* 4 (1953), 810–815.
2. C. C. Chang and H. J. Keisler, *Model Theory*, 3rd ed., North-Holland, Amsterdam, 1990.
3. R. Courant and H. Robbins, *What is Mathematics?: an elementary approach to ideas and methods*, 2nd ed. revised by Ian Stewart, Oxford University Press, New York, 1996.
4. G. Mohr, *Euclides Danicus*, Amsterdam, 1672; reprinted: Høst, Copenhagen, 1928.
5. J. Stillwell, *Elements of algebra: geometry, numbers, equations*, Springer, New York, 1994.
6. A. Tyszka, Some remarks on endomorphism rigid relations, *Univ. Iagel. Acta Math.* 31 (1994), 7–20.
7. L. Asimow and B. Roth, The rigidity of graphs, *Trans. Amer. Soc.* 245 (1978), 279–289.
8. H. Maehara, Distances in a rigid unit-distance graph in the plane, *Discrete Appl. Math.* 31 (1991), 193–200.
9. H. Maehara, Extending a flexible unit-bar framework to a rigid one, *Discrete Math.* 108 (1992), 167–174.

Centrum D 6 / 70
31-933 Kraków
Poland.
rttyszka@cyf-kr.edu.pl

PROBLEMS AND SOLUTIONS

Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**

with the collaboration of Paul T. Bateman, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Underwood Dudley, Michael A. Filaseta, Ira M. Gessel, Bart Goddard, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Robert Israel, Murray S. Klamkin, Daniel J. Kleitman, Fred Kochman, Frederick W. Luttmann, Frank B. Miles, Richard Pfiefer, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before March 31, 1998; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.*

PROBLEMS

10613. *Proposed by F. J. Flanigan, San Jose State University, San Jose, CA.* Fix a positive real number ν. Find all polynomials $P(x)$ with nonnegative real coefficients such that
(**a**) $P(0) = 0$, $P(1) = 1$, and $P(x) \le x^\nu$ for all $x \ge 0$.
(**b**) $P(0) = 0$, $P(1) = 1$, and $P(x) \ge x^\nu$ for all $x \ge 0$.

10614. *Proposed by Grigore-Raul Tataru, University of Bucharest, Romania.* Fix $p > 1$. Suppose that $a_1, a_2, \ldots$ is a sequence of positive real numbers such that $a_n a_{n+1} a_{n+2}^p + a_{n+2} - a_n = 0$ for all $n \ge 1$. Show that $\{a_n\}$ is convergent.

10615. *Proposed by Joaquin Gómez Rey, Alcorcón, Madrid, Spain.* For n a positive integer, evaluate

$$\sum (k_1 + k_2 + \cdots + k_n)! \prod_{i=1}^{n} \frac{i^{(i-1)k_i}}{(k_i!)(i!)^{k_i}}$$

where the summation runs over all n-tuples $(k_1, k_2, \ldots, k_n)$ of nonnegative integers such that $k_1 + 2k_2 + \cdots + nk_n = n$.

10616. *Proposed by Ernesto Bruno Cossi, Porto Alegra, Brazil.* Let K be a compact, convex set in the plane. For each interior point P of K and each line l through P, let A and B be the two points of l on the boundary of K, and let Q be the harmonic conjugate of P with respect to A and B. (That is, take Q to be collinear with A, P, and B so that $AP/PB = QA/QB$.) If K is an ellipse, then for each P the locus of points Q is a straight line. Is the converse true?

10617. *Proposed by James G. Merickel, Philadelphia, PA.* For a positive integer N, $\sigma(N)$ denotes the sum of the positive divisors of N. Given a positive integer n and a prime p, prove that there exist arbitrarily large sets S of multiples of n with the following property: For some positive integer m, the fraction $\sigma(N)/N$ reduces to a fraction whose denominator is p^m for every $N \in S$.

10618. *Proposed by S. Lakshminarayanan, S. L. Shah, and K. Nandakumar, University of Alberta, Edmonton, Canada.* Let A be a real $m \times n$ matrix of full rank with $m < n$ and let b be a real $m \times 1$ matrix. For $1 \leq i \leq n$, define

$$x_i = \frac{\det(A_i^* A^T) - \det(A_i A_i^T)}{\det(AA^T)},$$

where A_i^* is obtained by replacing the ith column of A by b and A_i is obtained by deleting the ith column of A. Show that $x = [x_1, \ldots, x_n]^T$ is a solution to the linear system $Ax = b$.

10619.* *Proposed by John Lawrence, Virginia Polytechnic Institute and State University, Blacksburg, VA.* Let X and Y be independent and identically distributed random variables in $\mathbb{R}^n$ with density $f(x)$ given by $g(||x||)$, where $||x||$ is the Euclidean norm and g is a strictly decreasing function (i.e., f is spherically symmetric about the origin and unimodal). For fixed $z \in \mathbb{R}^n$, let $h(z)$ denote the probability that the side joining X and Y is the longest side in a random triangle with vertices X, Y, and z. Must h be spherically symmetric about the origin and unimodal?

SOLUTIONS

Incircles of Curvilinear Triangles

10368 [1994, 273]. *Proposed by Emre Alkan, Bosphorus University, İstanbul, Turkey.* For each point O on diameter AB of a circle, perform the following construction. Let the perpendicular to AB at O meet the circle at point P. Inscribe circles in the figures bounded by the circle and the lines AB and OP. Let the R and S be the points at which the two incircles to the curvilinear triangles AOP and BOP are tangent to the diameter AB. Show that $\angle RPS$ is independent of the position of O.

Solution I by A. N. 't Woord, University of Technology, Eindhoven, The Netherlands. We prove that the line PR bisects $\angle APO$. Because the figure is symmetric, this implies that PS bisects $\angle OPB$. Hence $\angle RPS = \pi/4$.

A well known theorem of geometry says $OP^2 = AO \cdot OB$, where the expression XY denotes the length of the segment with endpoints X and Y. By adding the square of OB we obtain

$$BP^2 = OP^2 + OB^2 = AO \cdot OB + OB \cdot OB = AB \cdot OB. \qquad (1)$$

Let M be the center of the circle with diameter AB and F be the center of the circle touching AB, OP and the arc AP. Then $MF = AM - OR$. Square both sides to see that

$$RM^2 + OR^2 = MF^2 = (AM - OR)^2 = AM^2 - 2\,AM \cdot OR + OR^2.$$

Hence

$$AB \cdot BR - BR^2 = AR \cdot BR = (AM - RM)(AM + RM) =$$
$$= AM^2 - RM^2 = 2\,AM \cdot OR = AB \cdot OR,$$

and after rearrangement this yields

$$BR^2 = AB(BR - OR) = AB \cdot OB. \qquad (2)$$

Combining (1) and (2) gives $BR = BP$, so that $\angle PRB = \angle RPB$. This implies that $\angle RPO = \pi - \angle PRB = \pi - \angle RPB = \angle APR$.

Solution II by O. P. Lossers, University of Technology, Eindhoven, The Netherlands. The desired angle equals 45°, as we show in a slightly more general situation:

Proposition 1. *Extend both AB and PO to infinite lines. Then the given circle Γ, the line AB, and the line PO split the plane into eight curvilinear (in some cases infinite) triangles. Denote the points of contact of the "incircles" (circles that touch all three boundary lines) of these curvilinear triangles with the line AB by Q, R, S, and T, in that order with A between Q and R. Then $\angle QPR = \angle RPS = \angle SPT = 45^\circ$, PR bisects $\angle APO$, and PS bisects $\angle OPB$.*

Proof. Let O' denote the second point of intersection of PO and Γ. Let Δ denote the circle that has PO' as its diameter. The inversion with center P that maps O to O' transforms the configuration $\{\Gamma,\ AB,\ OO'\}$ into the configuration $\{l,\ \Delta,\ O'O\}$, where l is the diameter of Δ that is parallel to the tangent to Γ at P. This latter configuration has two perpendicular symmetry axes through O, and the incircles of the corresponding partitions have tangent points with Δ that halve the corresponding arcs of Δ. Since inversion in P preserves angles and takes lines through P to themselves, the results follow. □

This allows the following generalization.

Proposition 2. *Consider two circles meeting at a point P. Suppose that the common diameter of these circles meets one circle at point A and the other at O, where the segment AO meets the circles only at its endpoints. Inscribe a circle in the triangle whose sides are the line segment AO and the circular arcs AP and OP. Let the point where the inscribed circle touches AO be denoted by R. Then PR is the angle bisector of $\angle APO$*

Proof. Invert the configuration $\{l,\ \Delta,\ O'O\}$ of Proposition 1 about a point P of Δ not on $O'O$. □

Editorial comment. Many solvers used methods based on trigonometry or coordinates, and several included figures. The selected solutions are among the more purely geometric, and the description of the construction allows the reader to construct a figure easily.

Francisco Bellot Rosado included references to several other related problems, including MONTHLY problem 3887 [1938, 482; 1983, 486].

Solved also by P. M. Abe, M. Amengual (Spain), J. Anglesio (France), R. Barbara (Lebanon), F. Bellot Rosado (Spain), K. L. Bernstein, R. J. Chapman (U. K.), A. Coffman, R. G. Griswold, J.-P. Grivaux (France), R. Holzsager, D. J. Jones, H. Kappus (Switzerland), H. G. Killingbergtrø (Norway), P. G. Kirmser, N. Komanda, M. Lehtinen (Finland), H. M. Marston, A. Nijenhuis, C. G. Petalas (Greece), M. Reid, I. A. Sakmar (Turkey), J. Sarkar, B. Shawyer (Canada), D. Tang, J. Vlachos (Greece), N. D Vo (Canada), M. Vowe (Switzerland), H. Weingarten, R. L. Young, NSA Problems Group, and the proposer.

A Hidden Equilateral Triangle

10378 [1994, 363]. *Proposed by Bjorn Poonen, University of California, Berkeley, CA.* Given that point D is in the interior of $\triangle ABC$ and that there are real numbers a, b, c, d such that $AB = ab$, $AC = ac$, $AD = ad$, $BC = bc$, $BD = bd$, and $CD = cd$, prove that $\angle ABD + \angle ACD = \pi/3$.

Solution I by Jean Anglesio, Garches, France. We have

$$\frac{DB}{DC} = \frac{AB}{AC} = \frac{b}{c}, \quad \frac{DC}{DA} = \frac{BC}{BA} = \frac{c}{a}, \quad \text{and} \quad \frac{DA}{DB} = \frac{CA}{CB} = \frac{a}{b}.$$

Then D is one "isodynamic point" of the triangle ABC, that is, a point common to the three "Apollonius circles" of $\triangle ABC$ (see R. A. Johnson, *Modern Geometry*, Houghton Mifflin, 1929). We recall that one Apollonius circle of $\triangle ABC$ is the locus of points M such that $MB/MC = AB/AC$. The three Apollonius circles of $\triangle ABC$ have two points in common, called the *isodynamic points* of $\triangle ABC$. One of the two points is exterior to the circumcircle of $\triangle ABC$, Furthermore, if the other isodynamic point D is in the interior of $\triangle ABC$, then $\angle BDC = \angle BAC + \pi/3$, $\angle CDA = \angle CBA + \pi/3$, and $\angle ADB = \angle ACB + \pi/3$.

Now if E is the intersection of BD and AC, we have $\angle BDC = \angle DCA + \angle CED$ and $\angle CED = \angle BAC + \angle DBA$. Combining these, we obtain

$$\angle BAC + \pi/3 = \angle DCA + \angle BAC + \angle DBA,$$

from which the result follows.

Solution II by O. P. Lossers, University of Technology, Eindhoven, The Netherlands. Embedding the figure in the complex plane, so that the points A, B, C and D correspond to the complex numbers α, β, γ and δ, respectively, we observe that the three numbers $(\delta-\alpha)(\beta-\gamma)$, $(\delta-\beta)(\gamma-\alpha)$ and $(\delta-\gamma)(\alpha-\beta)$ all have the same modulus $abcd$. On the other hand, these three numbers add up to zero! Thus, they are the vertices of an equilateral triangle centered at the origin. In other words, we have

$$\frac{(\delta-\alpha)(\beta-\gamma)}{(\delta-\beta)(\gamma-\alpha)} = \frac{(\delta-\beta)(\gamma-\alpha)}{(\delta-\gamma)(\alpha-\beta)} = \frac{(\delta-\gamma)(\alpha-\beta)}{(\delta-\alpha)(\beta-\gamma)} = \exp\left(\pm\frac{2\pi i}{3}\right).$$

Hence,

$$\frac{\delta-\alpha}{\delta-\beta} = -\exp\left(\pm\frac{2\pi i}{3}\right)\frac{\gamma-\alpha}{\gamma-\beta} = \exp\left(\mp\frac{\pi i}{3}\right)\frac{\gamma-\alpha}{\gamma-\beta}.$$

Taking arguments, we get $\angle ADB = \angle ACB \pm \pi/3$. Since $\angle ADB > \angle ACB$ and the angle sum in quadrangle $ADBC$ equals 2π, the result follows.

Solution III by C. Kenneth Fan, Harvard University, Cambridge, MA. Construct three triangles $\triangle B_\alpha C_\alpha D_\alpha$, $\triangle A_\beta C_\beta D_\beta$, and $\triangle A_\gamma B_\gamma D_\gamma$, respectively similar to the three triangles $\triangle BCD$, $\triangle ACD$, and $\triangle ABD$, with similarity ratios a, b, and c (so that each of the new triangles has a side of length abc). Label each subscripted vertex to correspond to its unsubscripted counterpart.

By construction, $\angle B_\alpha D_\alpha C_\alpha + \angle A_\beta D_\beta C_\beta + \angle A_\gamma D_\gamma B_\gamma = 2\pi$. Furthermore, the lengths work out so that we may fit the constructed triangles together so that their interiors do not overlap and so that the edges $B_\alpha D_\alpha$ and $A_\beta D_\beta$ coincide, the edges $C_\beta D_\beta$ and $B_\gamma D_\gamma$ coincide, and the edges $A_\gamma D_\gamma$ and $C_\alpha D_\alpha$ coincide. We also require that the vertices D_α, D_β, and D_γ coincide. The result is an equilateral triangle of side length abc. Thus, $\angle ABD + \angle ACD = \angle A_\gamma B_\gamma D_\gamma + \angle A_\beta C_\beta D_\beta = \pi/3$.

Solution IV by Hans Georg Killingbergtrø, Horten, Norway. From the figure $ABCD$, make a similar figure $ACC'D'$, employing a scale of $c : b$ and $\angle CAB$ as angle of rotation about A, whereby $CD' = CD$ and $\angle DCD' = \angle ABD + \angle ACD$. Since $\angle D'AD = \angle CAB$ (the angle of rotation) and $AD'/AD = c/b = AC/AB$, $\triangle ADD'$ is similar to $\triangle ABC$ with a scale of $d : b$. Hence $DD' = BC \cdot d/b = cd = CD$, which shows that $\triangle CDD'$ is equilateral.

Editorial comment. The Apollonius circles also appear in H. S. M. Coxeter, *Introduction to Geometry*, Second edition, Wiley, 1969, section 6.6, p. 88, and D. Pedoe, *A Course of Geometry*, Cambridge, 1970, section 18.3, p. 77. In both of these references, this concept is used to introduce inversive geometry. In this spirit, one notices that solution II interprets the conclusion as a complex cross ratio, which is the fundamental invariant of inversive geometry. Indeed, inversive geometry leads to yet another solution, which we sketch here. First, the concept of *triangle* is not preserved by inversion, so it needs to be replaced. The line AB can be considered as the circle through A, B and the point at infinity. Thus, the image of the triangle under an inversive transformation can be recovered from the images of four points: A, B, C and the point at infinity. Meanwhile, each Apollonius circle is characterized as the points P for which a certain cross ratio of P with the points A, B, and C is purely imaginary. Thus, the isodynamic points depend only on the locations of A, B, C and not on the point playing the role of the point at infinity. In particular, if A, B, and C

are the vertices of an equilateral triangle, the two isodynamic points must be the center of the triangle and the point at infinity, and the result is clear. The general result follows since there is an inversive transformation taking any three points to the vertices of an equilateral triangle.

Another approach used Euler's formula for the volume of a tetrahedron (see Heinrich Dörrie, *100 Great Problems of Elementary Mathematics*, Dover, 1965, section 68) to express the condition that A, B, C, and D are coplanar, and the law of cosines to express the desired conclusion in terms of a, b, c, and d. It is then straightforward to relate the two conditions.

Solved also by R. Barbara (Lebanon), J. C. Binz (Switzerland), R. J. Chapman (U. K.), A. Coffman, S. B. Ekhad, M. S. Klamkin (Canada), J. H. Lindsey II, M. Reid, N. D Vo (Canada), M. Vowe (Switzerland), NSA Problems Group, and the proposer.

Some Volumes in Infinite Products of an Interval

10402 [1994, 682]. *Proposed by Werner Schindler, Universität Regensburg, Regensburg, Germany.* For every $j \in \mathbb{N}$ the term λ_j denotes a copy of the Lebesgue measure on the unit interval $[0, 1]$ while $\lambda^{\mathbb{N}}$ stands for the infinite product measure $\otimes_{j=0}^{\infty}\lambda_j$ on $[0, 1]^{\mathbb{N}}$. The mapping $f\colon [0, 1] \to [0, 1]$ is assumed to be measurable. Its n-fold application $f \circ f \circ \cdots \circ f$ is abbreviated as $f^{[n]}$ and $A(f) = \{\langle x_n\rangle \in [0, 1]^{\mathbb{N}} : x_n \le f^{[n]}(x_0) \text{ for each } n \ge 1\}$.
(**a**) Express $\lambda^{\mathbb{N}}(A(f))$ as a limit of one-dimensional definite integrals.
(**b**) Find a function f with $\lambda^{\mathbb{N}}(A(f)) = 1/2$.
(**c**) Compute $\lambda^{\mathbb{N}}(A(f))$ for $f(x) = e^{x-1}$.

Solution by A. N. 't Woord, University of Technology, Eindhoven, The Netherlands. (**a**) For $m \ge 1$, let

$$A_m(f) = \left\{\langle x_n\rangle \in [0, 1]^{\mathbb{N}} : x_n \le f^{[n]}(x_0) \text{ for } 1 \le n \le m\right\}.$$

Then

$$\lambda^{\mathbb{N}}\big(A_m(f)\big) = \int_0^1 f(x)f^{[2]}(x)\cdots f^{[m]}(x)\,dx.$$

Since $A(f) = \bigcap_{m=1}^{\infty} A_m(f)$, we get

$$\lambda^{\mathbb{N}}\big(A(f)\big) = \lim_{m\to\infty}\lambda^{\mathbb{N}}\big(A_m(f)\big) = \lim_{m\to\infty}\int_0^1 f(x)f^{[2]}(x)\cdots f^{[m]}(x)dx.$$

(**b**) Let $f(x) = \sqrt{x}$. Then $f(x)f^{[2]}(x)\cdots f^{[m]}(x) = x^{1-2^{-m}}$, so

$$\lambda^{\mathbb{N}}(A(f)) = \lim_{m\to\infty}\int_0^1 x^{1-2^{-m}}dx = \lim_{m\to\infty}\frac{1}{1-2^{-m}+1} = \frac{1}{2}.$$

(**c**) For $f(x) = e^{x-1}$ we have $\frac{df^{[m]}}{dx}(x) = f^{[m]}(x)\frac{df^{[m-1]}}{dx}(x)$, and so by induction $\frac{df^{[m]}}{dx}(x) = f(x)f^{[2]}(x)\cdots f^{[m]}(x)$. Since $e^x \ge 1 + x$ with equality only for $x = 0$, $f(x) \ge x$ with equality only for $x = 1$. Thus for $x \in [0, 1)$ the sequence $\langle f^{[n]}(x)\rangle$ is increasing and bounded, so it has a limit L. Continuity of f implies that L is a fixed point, so $L = 1$. Therefore,

$$\lambda^{\mathbb{N}}\big(A(f)\big) = \lim_{m\to\infty}\int_0^1 \frac{df^{[m]}}{dx}dx = \lim_{m\to\infty}\Big(f^{[m]}(1) - f^{[m]}(0)\Big) = 1 - 1 = 0.$$

Editorial comment. John H. Lindsey II gave a general form of the example of (**b**), using the same method to show that, for $f(x) = x^t$ with $0 \le t < 1$, $\lambda^{\mathbb{N}}(A(f)) = 1 - t$.

Solved also by C. Anderson, N. Bouzar, R. J. Chapman (U. K.), W. Hensgen (Germany), V. Hernández (Spain), R. Holzsager, J. H. Lindsey II, R. Stong, WMC Problems Group, and the proposer.

10409* [1994, 794]. *Proposed by J. van de Lune, Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands.* Let $p_1, p_2, p_3, \ldots$ denote the prime numbers in increasing order, and define $S_k = \{t \in \mathbb{R} : \sin(t \log(p_k)) > 0\}$, $C_k = \{t \in \mathbb{R} : \cos(t \log(p_k)) > 0\}$, $S_n^* = \cap_{k=1}^n S_k$, and $C_n^* = \cap_{k=1}^n C_k$. Prove (or disprove) that the relative measure of S_n^* and C_n^* (in $\mathbb{R}$) is equal to 2^{-n}. More precisely, prove (or disprove) that

$$\lim_{T\to\infty} \frac{1}{2T} \lambda\big(S_n^* \cap [-T, T]\big) = 2^{-n}$$

and the corresponding statement for C_n^*, where λ denotes Lebesgue measure.

Solution by National Security Agency Problems Group, Fort Meade, MD. First, we fix some notation and introduce a basic definition. For a subset J of $[0, 1]^n$, c_J denotes the indicator function of J; that is, $c_J(x) = 1$ or 0 depending on whether or not x is in J.

Let ω map $\mathbb{R}$ into $[0, 1]^n$. The function values $\omega(t)$ are said to be *C-uniformly distributed modulo* 1 if for every parallelepiped $J \subseteq [0, 1]^n$ we have

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T c_J(\omega(t)\, dt = \int_{[0,1]^n} c_J(\mathbf{x})\, d\mathbf{x}. \tag{1}$$

A theorem of Kronecker says: If $\alpha_1, \alpha_2, \ldots, \alpha_n$ are real numbers that are linearly independent over $\mathbb{Q}$, then the function values of $\phi(t) = (t\alpha_1, t\alpha_2 \ldots, t\alpha_n)$ with each component reduced modulo 1 are C-uniformly distributed modulo 1.

Kronecker's theorem applies to $\alpha_1, \alpha_2, \ldots, \alpha_n$, where $\alpha_i = \frac{1}{2\pi} \log p_i$ for distinct prime numbers $p_1, p_2, \ldots, p_n$, because if $\sum_{i=1}^n (a_i/b_i) \log p_i = 0$, then $\prod_{i=1}^n p_i^{a_i/b_i} = 1$, which implies $a_i/b_i = 0$ for all i.

With $\phi(t)$ so defined and with $J = (0, 1/2) \times (0, 1/2) \times \cdots \times (0, 1/2) \subseteq [0, 1]^n$, (1) yields

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T c_J\big(\phi(t)\big)\, dt = \int_{[0,1]^n} c_J(\mathbf{x})\, d\mathbf{x} = 2^{-n}. \tag{2}$$

Now, $c_J\big(\phi(t)\big) = 1$ if and only if there are integers m_i such that $2\pi m_i < t \log p_i < 2\pi m_i + \pi$ for $i = 1, 2, \ldots, n$. Put another way, $c_J\big(\phi(t)\big) = 1$ if and only if $\sin\big(t \log p_i\big) > 0$ for all i, i.e., $t \in S_n^*$. Therefore (2) yields

$$\lim_{T\to\infty} \frac{1}{T} \lambda\big(S_n^* \cap [0, T]\big) = \lim_{T\to\infty} \frac{1}{T} \int_0^T c_J\big(\phi(t)\big)\, dt = 2^{-n}. \tag{3}$$

A similar argument with J replaced by $-J = (1/2, 1) \times \cdots \times (1/2, 1) \subseteq [0, 1]^n$ shows that

$$\lim_{T\to\infty} \frac{1}{T} \lambda\big(S_n^* \cap [-T, 0]\big) = 2^{-n}, \tag{4}$$

and combining (3) and (4) yields the claim about the relative measure of S_n^*.

The case of C_n^* can be treated analogously. Here we replace $\phi(t)$ by $\psi(t) = \phi(t) + (1/4, \ldots, 1/4)$, where as usual the additions are modulo 1. By the translation invariance of λ, the values of $\psi(t)$ are also C-uniformly distributed modulo 1. With J as before, one finds $c_J\big(\psi(t)\big) = 1$ if and only if $\cos\big(\log p_i\big) > 0$ for all i. Now, (1) yields

$$\lim_{T\to\infty} \frac{1}{T} \lambda\big(C_n^* \cap [0, T]\big) = 2^{-n}. \tag{5}$$

Since the cosine is an even function, it is clear that $\lambda\big(C_n^* \cap [0, T]\big) = \lambda\big(C_n^* \cap [-T, 0]\big)$, which completes the proof of the claim about the relative measure of the C_n^*.

A reference for Kronecker's theorem and the basic properties of uniform distribution modulo 1 is Edmund Hlawka, *The Theory of Uniform Distribution*, A B Academic Publishers, Berkhamsted, U. K., 1984.

Editorial comment. Robin Chapman's solution followed the proof of Weyl's Equidistribution Theorem in T. W. Körner, *Fourier Analysis*, Cambridge, 1988.

Solved also by R. J. Chapman (U. K.), J. J. Dai, and J. Lauret (Argentina, n=2 only).

Divisors of the Falling Factorial

10431 [1995, 169]. *Proposed by Yury J. Ionin, Central Michigan University, Mt. Pleasant, MI.* For positive integers n and s with $n \geq s$, the *falling factorial* $(n)_s$ is defined as $n!/(n-s)!$. Let $d(n,s)$ denote the greatest common divisor of the falling factorials $(n)_s$ and $(n+s)_s$. Prove that $d(n,s) \mid (2s-1)_{\lfloor 4s/3 \rfloor}$.

Solution by A. N. 't Woord, University of Technology, Eindhoven, The Netherlands. For each prime p, let $v_p(x)$ be the largest integer k such that p^k divides x. It is sufficient to prove that $v_p(d(n,s)) \leq v_p((2s-1)_{\lfloor 4s/3 \rfloor})$ for every prime p. At most one integer in $\{n-s+1, \ldots, n+s\}$ has a divisor $p^k \geq 2s$. If such an integer exists and exceeds n, let $m = (n)_s$; otherwise, let $m = (n+s)_s$. In either case,

$$v_p\big(d(n,s)\big) \leq v_p(m) \leq \left\lceil \frac{s}{p} \right\rceil + \left\lceil \frac{s}{p^2} \right\rceil + \cdots + \left\lceil \frac{s}{p^t} \right\rceil,$$

where $p^t < 2s \leq p^{t+1}$. Also,

$$\begin{aligned} v_p\big((2s-1)_{\lfloor 4s/3 \rfloor}\big) &= \left(\left\lfloor \frac{2s-1}{p} \right\rfloor + \cdots + \left\lfloor \frac{2s-1}{p^t} \right\rfloor \right) \\ &\quad - \left(\left\lfloor \frac{2s - \lfloor 4s/3 \rfloor - 1}{p} \right\rfloor + \cdots + \left\lfloor \frac{2s - \lfloor 4s/3 \rfloor - 1}{p^t} \right\rfloor \right). \end{aligned}$$

Thus it is enough to prove that, for $a < 2s$,

$$L = \left\lfloor \frac{2s-1}{a} \right\rfloor - \left\lceil \frac{s}{a} \right\rceil \geq \left\lfloor \frac{2s - \lfloor 4s/3 \rfloor - 1}{a} \right\rfloor = R.$$

Write $s = qa + r$ with $q = \lfloor s/a \rfloor$. Then $L = q + \lfloor (2r-1)/a \rfloor - \lceil r/a \rceil$, and $R = q + \lfloor (r - 1 - \lfloor s/3 \rfloor)/a \rfloor$. For $0 \leq 2r \leq a$, we have $r - 1 - \lfloor s/3 \rfloor < 0$, so $3r \leq a + r \leq qa + r = s$. Consequently, $L = q - 1 \geq R$. For $2r > a$, we find $L = q \geq R$. Therefore, $L \geq R$ in all cases.

Solved also by J. H. Lindsey II, L. E. Mattics, and the proposer.

The Domain of Permissible Pairings of First and Second Moments

10505 [1996, 72]. *Proposed by Fu-Chen Chang, National Sun Yat-Sen University, Kaohsiung, Taiwan.* For $a, b \in \mathbb{R}$ with $a < b$ and $n \in \mathbb{N}$ with $n \geq 2$ let

$$S_n(a,b) = \left\{ (m_2, m_2) \ : \ m_1 = \frac{1}{n} \sum_{i=1}^{n} x_i, \ m_2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right\},$$

with $\{x_1, \ldots, x_n\}$ ranging over all possible samples of n numbers in the interval $[a, b]$. Find the area $\Delta_n(a,b)$ of $S_n(a,b)$.

Solution by Richard Holzsager, American University, Washington, DC. Shifting the interval does not change the area and stretching it by a factor r multiplies the area by r^3 (since the width of S_n gets multiplied by r and the height by r^2). Therefore, it suffices to look at a convenient interval and then deduce the general result. Choose $a = 0$, $b = n$. For any $0 \leq x \leq n$, if $m_1 = x$, then $m_2 = x^2 + \mathrm{V}ar(x_1, \ldots, x_n)$ varies from a minimum of x^2 to some maximum. The maximum occurs when the variance is greatest. This happens when all but at most one of the x_i's are at endpoints; this is easy to see, because if two are interior,

then raising the larger and lowering the smaller by the same amount increases the variance. So, if $i \le x \le i+1$, then 0 occurs $n-i-1$ times, n occurs i times, and the remaining term is $n(x-i)$. Then m_2 turns out to be $ni + n(x-i)^2$. Subtracting x^2 and integrating from i to $i+1$ gives $(n-1)/3 + (n-1)i - i^2$. Summing from $i = 0$ to $n-1$ and simplifying, we get $(n^3 - n^2)/6$. The answer for (a, b) is obtained by multipling by the cube of $(b-a)/n$, giving $(1 - 1/n)(b-a)^3/6$.

Editorial comment. Several solvers noted the geometrical formulation of this problem and its solution: Picture the cube $[a, b]^n$ in $\mathbb{R}^n$. (It is easiest to do this in three dimensions!) The value of m_1 is constant along hyperplanes perpendicular to the main diagonal of this cube (the one that passes through the origin). The value of m_2 is constant on spherical surfaces, because it is related to distance from the origin. The minimum value of m_2 on a hyperplane section within the cube is therefore on the diagonal, because this is the closest point to the origin; the maximum value is on an edge of the cube, since such points of a cross-section are farthest from the origin.

Now picture a hyperplane sweeping through the cube from nearest corner (to the origin) to farthest corner (from the origin), remaining always perpendicular to the main diagonal of the cube. At first, this hyperplane intersects the n edges emanating from the lower vertex; then at a certain point it contains a set of n vertices; then for a while it intersects certain edges till once again it contains a set of vertices; this pattern continues until the upper vertex is reached. Thus there are n episodes, each involving passing from intersection with a set of vertices through intersection with a set of edges to intersection again with a set of vertices. Within each episode the behavior of m_2 is parabolic. On the other hand, the behavior of m_1 is parabolic through the entire operation.

J. H. B. Kemperman referred to related results included in his paper: Moment problems for sampling without replacement, *Indagationes Mathematicae* **76** (1973) 181–188.

Solved also by R. A. Agnew, M. Benedicty, R. J. Chapman (U. K.), D. L. Farnsworth, M. Hoffman, J. H. B. Kemperman, M. S. Klamkin (Canada), J. H. Lindsey II, R. Richberg (Germany), K. Schilling, R. Stong, A. Tissier (France), T. V. Trif (Romania), A. N. 't Woord (The Netherlands), GCHQ Problems Group (U. K., two solutions), WMC Problems Group, USA Problems Group, and the proposer.

An Application of Crofton's Theorem

10512 [1996, 267]. *Proposed by V. A. Zalgaller, Steklov Mathematical Institute, Russian Academy of Sciences, St. Petersburg, Russia.* Let Q_1 and Q_2 be compact subsets of the Cartesian half-plane $y \ge 0$. Assume that both Q_1 and Q_2 contain points with $y > 0$. Let $\Phi_1 = \mathrm{Conv}(Q_1 \cup Q_2)$ and $l_1 = \mathrm{Len}(\partial\Phi_1)$. Let $I(Q)$ denote the set obtained by reflecting the set Q in the x-axis. Let $\Phi_2 = \mathrm{Conv}(Q_1 \cup I(Q_2))$ and $l_2 = \mathrm{Len}(\partial\Phi_2)$. Prove $l_2 > l_1$.

Solution by Sung Soo Kim, Hanyang University, Ansan, Kyunggi, Korea. Parameterize the lines in the plane by the pairs (r, θ), where $-\pi/2 < \theta \le \pi/2$ is the angle between the line and the horizontal, and r is the distance of the line from the origin. Crofton's Theorem (see for example L. A. Santaló, *Integral Geometry and Geometric Probability,* Addison-Wesley, 1976) says that for any compact convex set Q, the perimeter of Q is twice the area of the set of (r, θ) corresponding to lines that intersect Q. Let A be the set of lines that intersect Φ_1 and B be the set of lines that intersect Φ_2. We have to prove that B has greater area than A, or equivalently that $B - A$ has greater area than $A - B$.

Let $l \in A - B$. Then either (a) l is horizontal, intersects Q_2, and lies above Q_1, or (b) l is a slanted line with Q_1 and $I(Q_2)$ on one side, but Q_2 not entirely on that side. In either case, $I(l)$ is in $B - A$. Therefore $I(A - B) \subseteq B - A$. On the other hand, if l is horizontal, and just below the x-axis, then l is in $B - A$ and $I(l)$ is in B (and so not in $A - B$). In fact, the same is true of all lines in a sufficiently small neighborhood U of l, so $I(A - B) \subseteq B - A - U$. Since I preserves area, we are done.

Solved also by G. Pete (Hungary), R. Stong, the GCHQ Problem Group (U. K.), and the proposer.

A Regular Hexagon Emerging from a Triangle

10514 [1996, 267]. *Proposed by Jiro Fukuta, Shinsei-cho, Motosu-gun, Gifu-ken, Japan.* In $\triangle ABC$, Let P_1 and P_2, P_3 and P_4, P_5 and P_6 be the points on the sides BC, CA, AB respectively, such that

$$\frac{|BP_1|}{|P_1C|} = \frac{|CP_2|}{|P_2B|} = \frac{|CP_3|}{|P_3A|} = \frac{|AP_4|}{|P_4C|} = \frac{|AP_5|}{|P_5B|} = \frac{|BP_6|}{|P_6A|} = r$$

with $0 \le r \le 1$. Let A', B', C' be the points of intersection of P_1P_4 and P_2P_5, P_3P_6 and P_4P_1, P_5P_2 and P_6P_3, respectively. Let $Q_iP_iP_{i+1}$, $i = 1, \ldots, 6$ be the equilateral triangles built outwards on the sides of the hexagon $P_1P_2 \ldots P_6$. Let $R_iQ_{i-1}Q_{i+1}$, $i = 1, \ldots, 6$ be the equilateral triangles built outwards on the diagonals of the hexagon $Q_1Q_2 \ldots Q_6$.

(**a**) Show that the points Q_1, A', and Q_4 lie on R_1R_4.

(**b**) Show that the diagonals R_1R_4, R_2R_5, and R_4R_6 are concurrent and equal in length, and that the angle of intersection of any two of these lines is 60°.

(**c**) Let G_i be the centroid of the triangle $R_{i-1}R_iR_{i+1}$, $i = 1, \ldots, 6$. Show that $G_1 \ldots G_6$ is a regular hexagon and that its center coincides with the centroid of the triangle ABC.

Solution by Robin Chapman, University of Exeter, Exeter, U. K. Consider the triangle as embedded in the complex plane, and assume that A, B, and C appear in that order as the perimeter is traversed in an anticlockwise direction. Let the complex numbers $(r+1)\alpha$, $(r+1)\beta$, and $(r+1)\gamma$ correspond to A, B, and C, respectively. Then $P_1 = \beta + r\gamma$, $P_2 = r\beta + \gamma$, $P_3 = \gamma + r\alpha$, $P_4 = r\gamma + \alpha$, $P_5 = \alpha + r\beta$, $P_6 = r\alpha + \beta$. The quadrilateral $AP_4A'P_5$ is a parallelogram and so $A' = P_4 + P_5 - A = (1-r)\alpha + r\beta + r\gamma$. Note that if XYZ is an equilateral triangle labelled anticlockwise, then $Z = \overline{\zeta}X + \zeta Y$ where $\zeta = \exp(\pi i/3) = (1+\sqrt{-3})/2$. Hence $Q_i = \zeta P_i + \overline{\zeta} P_{i+1}$ and $R_i = \zeta Q_{i-1} + \overline{\zeta} Q_{i+1}$. In particular,

$$Q_1 = (\zeta + r\overline{\zeta})\beta + (r\zeta + \overline{\zeta})\gamma, \quad Q_4 = \alpha + r\overline{\zeta}\beta + r\zeta\gamma,$$
$$R_1 = -r\alpha + (\zeta + r)\beta + (\overline{\zeta} + r)\gamma, \quad \text{and} \quad R_4 = (2-r)\alpha + (\overline{\zeta}^2 + r)\beta + (\zeta^2 + r)\gamma.$$

It follows that $R_4 - R_1 = 2(\alpha + \overline{\zeta}^2\beta + \zeta^2\gamma)$. Similarly, $R_6 - R_3 = 2(\beta + \overline{\zeta}^2\gamma + \zeta^2\alpha) = \zeta^2(R_4 - R_1)$ and $R_2 - R_5 = 2(\gamma + \overline{\zeta}^2\alpha + \zeta^2\beta) = \overline{\zeta}^2(R_4 - R_1)$. Hence the line segments R_1R_4, R_3R_6, and R_5R_2 all have the same length and make angles $\pi/3$ with each other. Also $A' = rQ_1 + (1-r)Q_4$, $R_1 = (1+r)Q_1 - rQ_4$, and $R_4 = (r-1)Q_1 + (2-r)Q_4$, so Q_1, Q_4, A', R_1, and R_4 are all collinear.

Now $3G_i = R_{i-1} + R_i + R_{i+1}$; in particular $3G_1 = (r-1)\alpha + (r+1+2\zeta)\beta + (r+1+2\overline{\zeta})\gamma$ and $3G_2 = (2\overline{\zeta} - 1 + r)\alpha + (2\zeta - 1 + r)\beta + (3+r)\gamma$. If H is the centroid of $\triangle ABC$ then $3H = (r+1)(\alpha + \beta + \gamma)$. It follows that $G_1 - H = (2/3)(-\alpha + \zeta\beta + \overline{\zeta}\gamma)$ and $G_2 - H = (2/3)(\overline{\zeta}^2\alpha + \zeta^2\beta + \gamma) = \zeta(G_1 - H)$. Extending this computation gives $G_i - H = \zeta^{i-1}(G_1 - H)$ for all i, and so the G_i are the vertices of a regular hexagon with centre H.

It remains to show only that the lines R_1R_4, R_2R_5, and R_3R_6 are concurrent. The points R_5, R_1, and R_3 are the third vertices of the equilateral triangles constructed outwards on the sides of $\triangle Q_2Q_4Q_6$. The lines Q_2R_5, Q_4R_1 and Q_6R_3 are then concurrent in the *Fermat point* of $\triangle Q_2Q_4Q_6$ (see §4.2 of H. S. M. Coxeter & S. L. Greitzer, *Geometry Revisited,* Mathematical Association of America, 1967). But since Q_2 lies on the line R_2R_5, the line R_2R_5 is the line Q_2R_5 and so on. Hence the lines R_2R_5, R_4R_1, and R_6R_3 are concurrent.

Solved also by M. Benedicty, Z. Čerin (Croatia), J. Duncan, N. Komanda, O. P. Lossers (The Netherlands), G. Pete (Hungary), R. Simon (Chile), T. Trif (Romania), R. Young, the Con Amore Problem Group (Denmark), and the proposer.

10518 [1996, 347]. *Proposed by Yunan Diao, Kennesaw State College, Marietta, GA.* Let X be a finite set of points in a metric space and let X_1 and X_2 be a partition of X into two disjoint nonempty subsets. Let $d(X_1, X_2) = \min\{d(x_1, x_2) : x_1 \in X_1, x_2 \in X_2\}$ be called the distance between the subsets, and let the largest value of the distance between two such subsets be called the *splitting number* of X.

If X consists of n random points, independently selected from the uniform distribution on a ball of radius 1 in 3-dimensional Euclidean space, show that the splitting number of X is almost surely small. More precisely, for $a < 1$, show that there is a constant $\alpha > 0$ depending only on a such that the splitting number of X is less than a with probability at least $1 - e^{-\alpha n}$.

Solution by Jonathan Pillow, University of Arizona, Tucson, AZ. Let S_n denote the splitting number of the n points in the set X. Given $a < 1$, create a 3-dimensional cubic grid, spaced $a/\sqrt{6}$ on a side, and intersect it with a ball of radius 1 in $\mathbb{R}^3$, partitioning the ball into m open regions, $A_1, A_2, \ldots, A_m$, and some boundary points (the grid planes and the surface of the sphere). Note that, if the ball is placed tangent to a grid plane in each direction, then $m \leq \lceil 2\sqrt{6}/a \rceil^3$. Also note that, for any two points x and y in regions A_i and A_j that are adjacent (along a face), we have $d(x, y) < a$.

The probability that X contains a boundary point is 0, so assume that $X \subseteq \bigcup_{i=1}^m A_i$. If $A_i \cap X$ is nonempty for all regions A_i, $1 \leq i \leq m$, then we must have $S_n < a$, because for every nontrivial partition of X into sets X_1 and X_2, there must exist $x \in X_1$ and $y \in X_2$ such that x and y are in adjacent regions, implying $d(X_1, X_2) < a$. Thus, $S_n \geq a$ implies that at least one region A_i contains no point of X, so

$$P(S_N \geq a) \leq P(\exists i\ A_i \cap X = \emptyset) \leq \sum_{i=1}^{m} P(A_i \cap X = \emptyset).$$

Now, the probability that A_i contains no points of X is $(1 - 3|A_i|/4\pi)^n$, where $|A_i|$ is the volume of A_i. Taking A_1 to be the smallest A_i, we get $P(S_n \geq a) \leq m(1 - 3|A_1|/4\pi)^n$.

We now argue that $P(S_n \geq a) \leq e^{-\alpha n}$, where α is derived in the following manner. Observe that, if $n \geq n_0 = (4\pi/|A_1|)\log m$, then

$$\begin{aligned}\log P(S_n \geq a) &\leq \log m + n \log\left(1 - \frac{3|A_1|}{4\pi}\right) \\ &\leq \log m - \frac{3n|A_1|}{4\pi} = -n\left(\frac{3|A_1|}{4\pi} - \frac{\log m}{n}\right) \leq -n\left(\frac{2|A_1|}{4\pi}\right) = -\alpha_1 n,\end{aligned}$$

so $P(S_n \geq a) \leq e^{-\alpha_1 n}$. On the other hand, if $n < n_0$, note that $P(S_n \geq a) \leq e^{n(\log P_0)/n_0} = e^{-\alpha_2 n}$, where $P_0 = \max_{n<n_0} P(S_n \geq a)$. It follows that $P(S_n < a) \geq 1 - e^{-\alpha n}$ for all n, where $\alpha = \min(\alpha_1, \alpha_2)$.

Solver's note: Better estimates for α may be obtained if greater care is used in partitioning the ball into regions. If we consider points randomly distributed in $[0, 1]^3$, then the analysis is much simpler and we can easily show that $\alpha \geq a^3/2\sqrt{6}$. We may also let a vary with n. Jessica Sidman of Scripps College and I considered such questions at the Michigan Tech REU program, and we proved that, for points chosen from $[0, 1]^2$,

$$\lim_{n\to\infty} P\left(\sqrt{\frac{\log n}{\pi n}} \leq S_n \leq \sqrt{\frac{5\log n}{2n}}\right) = 1.$$

Solved also by J. H. Lindsey II, R. Stong, and the proposer.

REVIEWS

Edited by **Underwood Dudley**
Mathematics Department, De Pauw University, Greencastle, IN 46135

An Introduction to Difference Equations, by Saber Elyadi. Springer-Verlag, New York, 1996, 336 pp, $45.

Reviewed by **Ronald E. Mickens**

The Newtonian revolution in physics led to the use of differential equations as the fundamental basis for the mathematical modeling of dynamical systems. Paradoxically, the great success of this paradigm also led to the widespread use of difference equations as discrete models of differential equations, primarily for the purposes of numerical integration of these equations. More recently, the easy access to digital computers provided a strong incentive for their users to have a knowledge of at least the elementary properties of difference equations. Modern books on the subject began to appear in the 1950's. One of the first and most popular of them was written by Goldberg [**3**]. It gave a general introduction to linear difference equations and applied them to the formulation and solution of problems in economics, psychology, and sociology. During the decade of the 1960's several books appeared for the mathematically more sophisticated reader, for example those by Levy and Lessman [**10**], Brand [**2**], Hildebrand [**4**], and Miller [**14**].

The past ten years has seen the arrival of books that go beyond just giving an introduction to the various elementary properties of difference equations. Collectively [**1**, **5**, **6**, **7**, **8**, **9**, **12**, **13**, **15**], they consider the generalization to difference equations of many of the techniques usually associated with linear and nonlinear differential equations. They raise questions, and provide (partial) answers, to questions such as

- How can the asymptotic behavior of the solutions be determined?
- For difference equations with a nonlinear term multiplied by a "small" parameter, can perturbation methods be devised to provide uniform approximations to the solution?
- For difference equations, what corresponds to the Hopf bifurcation theorem?
- Can "exact" difference equation models of differential equations be constructed?
- How can accurate numerical solutions be determined computationally for all the solutions of a given linear difference equation?

The recent blossoming of research in difference equations and their application to problems in nonlinear discrete dynamics led Gerry Ladas and Saber Elaydi to found the *Journal of Difference Equations and Applications* (published by Gordon and Breach), whose first issue was published in January 1995. A major reason for this new journal was the need for a place where researchers in difference

equations could publish excellent papers on their results and have them read by colleagues in the field. Another factor was the lack of any mathematics journal that emphasized work in difference equations. In the past, researchers in this area usually published papers in journals devoted mainly to differential equations or in journals dealing with specific discipline-based applications.

Much of the current interest in difference equations within the general scientific community had its genesis in a review paper written by Robert May in 1976 **[11]**. He did an analytical and computational investigation of the properties of the solutions to

$$x_{k+1} = \lambda x_k(1 - x_k), \quad \lambda \geq 1. \tag{1}$$

This equation can be considered a discrete model for the growth of a single population using a finite-difference approximation to the so-called logistic differential equation

$$\frac{dy}{dt} = y(1 - y). \tag{2}$$

That is, if y_k is an approximation to $y(t_k)$, where $t_k = hk$ with $\Delta t = h$, then using a forward Euler approximation for the derivative gives the scheme

$$\frac{y_{k+1} - y_k}{h} = y_k(1 - y_k).$$

The substitution

$$x_k = \left(\frac{h}{1 + h}\right) y_k, \quad \lambda = 1 + h,$$

leads to (1). Now for positive initial data, $y(0) = y_0 < 0$, all the solutions of (2) go monotonically to the value $y = 1$. However, it is rather easy to demonstrate, as May **[11]** does, that (1) exhibits a variety of solution behaviors (periodic, chaotic, etc.) depending on the value selected for λ.

While (1) has a simple mathematical form, it turns out that many phenomena in the sciences can be modeled by simple difference equations whose solutions describe complex dynamical behavior **[8**, **9**, **15]**. A major advantage of discrete models for such systems is that, while the equations of motion cannot, in general, be solved analytically in terms of elementary functions, the solutions can be determined easily with the aid of digital computers. The results can then be displayed in a variety of visual formats for study and analysis. These facts have led many colleges and universities to create courses on discrete dynamical systems and/or difference equations. One consequence of these new courses was the publication of several books that could be used as introductions to discrete dynamical systems and difference equations or to more advanced topics in the theory of difference equations. The book under review combines both of these features.

The author, Saber Elyadi, is Professor of Mathematics at Trinity University in San Antonio, Texas. He has made important contributions in difference equations, especially in the areas of the discrete Levinson's theorem and the theory of Volterra difference equations. His book consists of eight chapters and is based on a course that he teaches at Trinity. The students in the course are upper-level undergraduates and come largely from mathematics and the physical and engineering sciences. The background required of the students is rather minimal, calculus and linear algebra, but certain topics from advanced calculus are needed for material near the end of the book.

The first three chapters introduce the reader to the fundamental concepts needed to understand both linear and nonlinear difference equations. In particular, the author does an excellent job in his presentation of the criteria for the asymptotic stability of fixed points (Chapter 1) and in its generalization in Chapter 4, where he discusses determination of stability by both linear approximation and Liapunov's second method.

Chapter 5 gives a through discussion of the Z-transform method, which serves exactly the same function for difference equations as the Laplace transform does for differential equations. In addition to showing how this technique can be used to solve linear difference equations, the author applies it to the scalar case of Volterra difference equations of convolution type.

Chapter 6 is on control theory. The chapter considers only time-invariant (autonomous) discrete systems and covers the basic concepts needed for an introduction to this topic: controllability, observability, and stabilizability by feedback. Chapter 7 gives an excellent introduction to various techniques that can be applied to determine the asymptotic behavior of solutions to both linear and nonlinear difference equations. In addition to discussing the well-known theorems of Poincaré and Perron, the author includes some of his own recently discovered results in this area. The final Chapter 8 is on oscillation theory. Let a nontrivial solution to a difference equation be denoted by x_k. The solution is said to be *oscillatory* (around zero) if for every positive integer N there exists $k \geq N$ such that $x_k x_{k+1} \leq 0$. If this is not the case, the solution is said to be *non-oscillatory*. The author presents an introduction to this topic and includes references where a more advanced treatment can be found.

There are several features that I especially like about the book. First, it contains a very extensive set of exercises at the end of each section. They are used not only as applications of the previously given theory but also, in many cases, help the reader extend the theoretical discussion given earlier in the section. Second, many applications are given for a variety of disciplines. For example, the materials in Sections 3.5.1–3.5.3 on Markov chains are excellent. Third, the book also includes several programs written especially for the TI-85 calculator. This feature will certainly help the reader discover various interesting features of difference equations through actual experimentation with the calculator.

The book does contain some typos and misprints. However, the alert reader can locate them rather easily. (My copy of the book came with an errata sheet.) I found the book to be well-written and very suitable for a good solid introduction to the fundamentals of difference equations, some of their applications, and several related advanced topics.

REFERENCES

1. Agaerwal, R. P., *Difference Equations and Inequalities*, Marcel Dekker, 1992.
2. Brand, L., *Differential and Difference Equations*, Wiley, 1966.
3. Goldberg, S., *Introduction to Difference Equations*, Wiley, 1958.
4. Hildebrand, F. B., *Finite-Difference Equations and Simulations*, Prentice-Hall, 1968.
5. Jerri, A. J., *Linear Difference Equations with Discrete Transform Methods*, Kluwer, 1996.
6. Immink, G. K., *Asymptotics of Analytic Difference Equations*, Springer-Verlag, 1983.
7. Kelley, W. G. and A. C. Peterson, *Difference Equations*, Academic Press, 1991.
8. Kocic, V. L. and G. Ladas, *Global Behavior of Nonlinear Difference Equations of Higher Order and Applications*, Kluwer, 1993.
9. Lakshmikantham, V. and D. Trigante, *Theory of Difference Equations*, Academic Press, 1988.
10. Levy, H. and F. Lessman, *Finite Difference Equations*, Macmillan, 1961.
11. May, R. M., Simple mathematical models with very complicated behavior, *Nature* **261** (1976)

459–467.
12. Mickens, R. E., *Difference Equations*, Van Nostrand Reinhold, 1987.
13. Mickens, R. E., *Difference Equations: Theory and Applications*, Chapman and Hall, 1990.
14. Sandefur, J. E., *Discrete Dynamical Systems*, Oxford, 1990.
15. Wimp, J., *Computation and Recurrence Relations*, Pitman, 1984.

Department of Physics
Clark Atlanta University
Atlanta, GA 30314
rohrs@math.gatech.edu

Calculus Lite, by Frank Morgan. A K Peters, Wellesley, 1995, 281 pp., $29.95.

Reviewed by **Wayne Roberts**

The calculus reform movement, when it started, had "lean and lively" as its slogan. This book makes no claim to having been influenced by the themes of reform, and there is considerable evidence to justify the silence. Nevertheless, with respect to the elusive goal of being lean this book, only 281 pages from cover to cover with each page only 6″ by 9″, succeeds where other books inspired by the reform movement have conspicuously failed. I have observed at several conferences on calculus reform that if mathematicians had but looked at themselves, they might have known in the beginning that it is easier to be lively than to be lean. Difficult though it may be to write a leaner calculus, however, it does seem worthwhile to write a little more about the virtue we hoped to achieve by writing less.

Stated succinctly, the goal was to focus attention on the key ideas of calculus. It was thought that calculus books exceeding one thousand pages in length were obscuring the structure of the calculus edifice by dwelling too much on the surrounding appurtenances. Where Gauss was concerned lest a fine building be obscured by scaffolding, the would-be reformers were concerned that it might be obscured by bric-a-brac.

Let us now ask what has actually happened. Some books have gotten smaller. Those that have made the most visible progress have done it, however, by the simple expedient of writing a one-year single-variable text, leaving the multivariable material to another book. This somehow misses the spirit of the thing. It is true, of course, that many of the newer books do omit topics commonly found in their weightier predecessors, but since shorter was not for the sake of shortness but for the sake of emphasis on key ideas, we are compelled to ask not if the newer books are shorter, but whether they really do focus student attention on the main ideas. I am not sure the answer is as clear as we would wish.

The readers of these newer books are being directed to the right ideas. The derivative of a function at a point is the slope of the line tangent to the graph of the function at that point. The integral of a function over an interval is the area under the graph of the function. The miracle is that these two ideas are related in a way that, from the right point of view, seems natural. I doubt, however, that even the best of the readers are yet seeing these truths jumping out from their texts. Instead of being obscured by more and more material, they are obscured by more

and more lengthy explanations. In this respect, Morgan seems closer to the original spirit by proclaiming his intention of "getting right to the point, and stopping there."

Morgan also illustrates, in his handling of the role of continuity in maxima and minima problems, how to avoid a tendency of authors that has been likened to giving a youngster a first hammer, and then instead of illustrating its usefulness in pounding nails, going immediately to a warning that it is not to be used for driving screws. After discussing a number of situations in which intuition is a reliable guide, he quotes a theorem: *A continuous function $y = f(x)$ on a closed interval $[a, b]$ attains a maximum and a minimum.* He then acknowledges that "it sounds right and obvious that a function will have a maximum and a minimum somewhere, and the details about 'continuous' and 'on a closed interval $[a, b]$' do not sound important." There follows a short discussion (half a page) that draws on some pictures to illustrate some possible troubles and concludes with, "The proof of this delicate theorem depends on a deep understanding of real numbers, continuous functions, and the difference between closed and open intervals. You can learn about it in a course in real analysis."

It is possible to show how something works when everything is going right, to be honest about the fact that sometimes things go wrong, to indicate that difficulties can be adequately dealt with, and still be concise. It is regrettable that authors reaching for the laudable goals of making things seem intuitive and natural so often feel that the goal precludes stating an important definition or theorem crisply, or saying anything about the necessity of ultimately undertaking a more careful analysis. It need not be so.

There are numerous ways in which Morgan achieves his brevity. He proves that the derivative of t^n is nt^{n-1} for $n = 1, 2, 3$ and then for any integer n, all on one page, and shortly (sixteen pages) has the student using the sum, product, quotient, and chain rules for differentiation of expressions built from polynomials. He reviews all one needs to know about sines and cosines, and derives in a highly intuitive way their derivatives, all in eight pages. He restricts discussion of finding antiderivatives to the use of substitution (four pages), tables (four pages), parts (one and one-half pages) and partial fractions (three pages), about which he says that "Focusing ... on distinct, linear factors provides the theory and ninety percent of the applications without the time-consuming algebra."

At times he abandons any appeal to intuition, resorting to what might best be described as blurting out some rules, as he does, for example, in his treatment of the laws of logarithms. His unmotivated definition of e as "the bank balance after one year of one dollar invested at an interest rate of 100% compounded continuously" misses a wonderful opportunity to let students discover an important number for themselves, and the quick move to $(a^x)' = (\ln a)a^x$ is a deplorable example of derivation by decree.

Letting students discover e for themselves is not an option for Morgan, however, determined as he is to write a text that in no way acknowledges the existence of calculators. Indeed, it is not until page 193 when, at the close of his four-page chapter on numerical methods, he acknowledges that computers exist. There certainly are materials written that contrive problems with no apparent purpose except to drive the student to technology to help with complex computations. If this were the only use to be made of such help, one might agree with Morgan's apparent conclusion that it's better to focus on ideas to the exclusion of any use of calculators or computers. There are better ways to use the tools now available, however. Students can discover things they once had to be told (they can,

with a calculator and the definition of the derivative, discover the number b for which b^x will be its own derivative). They can understand important ideas once passed over as abstractions that weren't likely to be on the test anyway (they can, with the use of the integrate key on their calculator, make a table of values for $f(x) = \int_1^x \sqrt{t^2 + 4}\, dt$ and so come to understand that an integral with a variable limit really does define a function). They can work on intrinsically interesting problems that don't have to be avoided because of computational complexity. To completely ignore this potential in a book with a 1995 copyright, as Morgan does, seems curious at best.

There are numerous other ways in which Morgan makes it clear that he is not, with his lean book, trying to cozy up to the reformers. There is nothing in his choice of applications that yields to the call to deal with topics that students will see as interesting, relevant, or exciting; he is still fencing animals in rectangular pens. The exercise lists are very much of the type that can be demolished at a keyboard while little learning takes place. There is no concession to the idea of including some open-ended questions that invite alternative approaches, that lend themselves to extended investigation over time, that offer the opportunity to exercise imagination, or that demand a coherent defense of the methodology chosen or the answer obtained.

Frank Morgan has demonstrated that a lean book can be written in a way that highlights key ideas. His clear delineation of differentiation, antidifferentiation, the integral, and the place of the fundamental theorem is exemplary in this regard. That is a contribution toward the goals the reformers had in mind. This reviewer, for one, wished he had put his gift for clear exposition in the employ of other goals of the reform movement.

Provost's Office
Macalester College
St. Paul, MN 55105
robertsw@macalstr.edu

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1–4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

Mathematics Appreciation, P, L*. *Mathematics: The Science of Patterns: The Search for Order in Life, Mind, and the Universe.* Keith Devlin. Scientific American Library, 1997, vii + 216 pp, $19.95 (P). [ISBN 0-7167-6022-3; 0-7167-5047-3] Updated to reflect recent proof of Fermat's last theorem. Incorporates minor corrections and a few additions to list of books for further reading (TR, April 1995). AO

Recreational Mathematics, S, L. *Which Way Did the Bicycle Go? ... and Other Intriguing Mathematical Mysteries.* Joseph D.E. Konhauser, Dan Velleman, Stan Wagon. Dolciani Math. Expos., No. 18. MAA, 1996, xv + 235 pp, $24.95 (P). [ISBN 0-88385-325-6] Almost two-hundred problems selected from 25 years of Macalester College's "Problem of the Week." Most of the problems should be accessible to (if not always solvable by) high school students. Solutions and references. Lots of fun. JO

Recreational Mathematics, S. *A Mathematical Jamboree.* Brian Bolt. Cambridge Univ Pr, 1995, 111 pp, $17.95 (P). [ISBN 0-521-48589-4] 114 miscellaneous mathematical puzzles, some new and some old chestnuts. Most are appropriate for high school students, but a few are harder. Includes solutions and instructions for building several kinds of harmonographs (Spirograph™-like devices). JO

Recreational Mathematics, S. *Winning Solutions.* Edward Lozansky, Cecil Rousseau. Problem Books in Math. Springer-Verlag, 1996, x + 244 pp, $34.95 (P). [ISBN 0-387-94743-4] Intended to bridge the gap between the mathematics typically taught in high schools and the mathematics required for success in high-level mathematical competitions. Sections on number theory, inequalities, and combinatorics. JO

Recreational Mathematics, P. *The Red Book of Mathematical Problems.* Kenneth S. Williams, Kenneth Hardy. Dover, 1996, ix + 174 pp, $6.95 (P). [ISBN 0-486-69415-1] Republication, with corrections, of *The Red Book: 100 Practice Problems for Undergraduate Mathematics Competitions* published by Integer Press in 1988 (TR, February 1989).

Precalculus, T(13: 1). *Workshop Calculus: Guided Exploration with Review, Volume 1.* Nancy Baxter Hastings, *et al.* Springer-Verlag, 1997, xxii + 391 pp, $29.95 (P). [ISBN 0-387-94611-X] Laboratory workbook. Integrates precalculus with topics from calculus including limits, the derivative, and the definite integral. Stresses individual and group discovery using a computer algebra system. PG

Precalculus, T(13). *Algebra with Applications.* William J. Adams. Kendall/Hunt, 1995, xii + 361 pp, $52.44 (P). [ISBN 0-7872-0996-1] Curiously old-fashioned treatment with many "hows" and few "whys." Pre-modern math era treatment uses "transposition" (transfer a term from one side of an equation to another and change its sign) as a basic operation for solving equations and inequalities. Similar "mindless rule" approach to most topics. Of questionable value for any purpose. MW

Education, S(15-17). *Ideas: NCTM Standards-Based Instruction.* Ed: Michael C. Hynes. NCTM, $11.50 (P) each. *Grades K–4*, 1995,

119 pp, [ISBN 0-87353-422-0]; *Grades 5–8*, 1996, v + 129 pp. [ISBN 0-87353-426-3] Each book contains nearly 50 reproducible one-page activity sheets adapted from the "Ideas" department in the *Arithmetic Teacher*. Good source of activities for pre-service and in-service mathematics teachers. MW

Education, T(14–16: 2). *Mathematics for Elementary Teachers: An Interactive Approach, Second Edition.* Thomas Sonnabend. Saunders College, 1997, xxii + 928 pp, $51.50. [ISBN 0-03-018367-7] Straightforward presentation of standard topics with interspersed "lesson exercises" and frequent calls for students to explain or predict outcomes. This edition updates suggested readings and technology coverage (though technology is still not integral). (*First Edition*, TR, May 1994.) MW

Education, P. *Towards Gender Equity in Mathematics Education: An ICMI Study.* Ed: Gila Hanna. ICMI Stud. Ser., V. 3. Kluwer Academic, 1996, xii + 304 pp, $133. [ISBN 0-7923-3921-5] Articles explore issues raised at 1993 International Commission on Mathematical Instruction Conference. Describes 13 national education systems from the perspective of gender equity, opportunity, classroom experiences, and educational outcomes. Research summaries; cross-cultural perspectives. MW

Education, S(17–18), P. *A Cognitive Analysis of U.S. and Chinese Students' Mathematical Performance on Tasks Involving Computation, Simple Problem Solving, and Complex Problem Solving.* Jinfa Cai. Journal for Res. in Math. Educ., Mono. No. 7. NCTM, vii + 151 pp, $7.50 (P). [ISBN 0-87353-424-7] Analyzes performance of U.S. and Chinese 6th grade students on different types of tasks. Chinese students substantially outperformed U.S. students on computation tasks, but not on complex problem-solving tasks. Surprisingly, Chinese students had much higher non-response rates on open-ended problems. MW

Logic, T(15–16: 1, 2), L. *Set Theory, Logic and their Limitations.* Moshé Machover. Cambridge Univ Pr, 1996, ix + 288 pp, $24.95 (P); $60. [ISBN 0-521-47998-3; 0-521-47493-0] Lecture notes for philosophy and mathematics students. Covers axiomatic set theory, cardinals and ordinals, propositional and first-order logic, and limitative results of Skolem, Tarski, Church, and Gödel. Concise presentation with few examples. Good remarks on history and philosophical issues. KES

Logic, P. *Henkin-Keisler Models.* George Weaver. Math. & Its Applic., V. 392. Kluwer Academic, 1997, xii + 253 pp, $146. [ISBN 0-7923-4366-2] Henkin used the method of constants to construct interpretations for first-order languages. Keisler modified the method to motivate the ultraproduct construction. Monograph describes Keisler's method and illustrates its use as an alternative to ultraproducts. KES

Foundations, T(14: 1), L. *A Transition to Advanced Mathematics, Fourth Edition.* Douglas Smith, Maurice Eggen, Richard St. Andre. Brooks/Cole, 1997, viii + 344 pp, $62.95. [ISBN 0-534-34028-8] Discusses methods of mathematical proof. Covers material on sets, relations, functions, cardinality, groups, and the completeness of the real numbers. (*Third Edition*, TR, November 1990; Extended Review, February 1991.) PG

Combinatorics, P. *Sperner Theory.* Konrad Engel. Ency. of Math. & Its Applic., V. 65. Cambridge Univ Pr, 1997, ix + 417 pp, $69.96. [ISBN 0-521-45206-6] Starts with statement and proof of Sperner's theorem. Develops theory of extremal problems on finite partially ordered sets. Covers flow-theoretic approach in Sperner theory; matchings and symmetric chain orders; probabilistic and algebraic methods; Macaulay posets. LC

Combinatorics, T(17), P. *One-Factorizations.* W.D. Wallis. Math. & Its Applic., V. 390. Kluwer Academic, 1997, xiv + 242 pp, $143. [ISBN 0-7923-4323-9] A one-factorization of a graph is a decomposition of the edge set of the graph into edge disjoint factors (spanning graphs) which are regular of degree 1. This text covers one-factors and one-factorizations, their applications, and connections to design theory. Includes exercises, extensive references. LC

Combinatorics, T(15–17: 1, 2), L. *Graphs & Digraphs, Third Edition.* Gary Chartrand, Linda Lesniak. Chapman & Hall, 1996, x + 422 pp, $59.95. [ISBN 0-412-98721-X] Changes from *Second Edition* (TR, December 1986): expanded treatment of domination in graphs, Hamiltonian graphs, graph decompositions, extremal graph theory; introductions to voltage graphs, graph labelings, and probabilistic graph theory; additional exercises; updated bibliography. KES

Combinatorics, T(17-18: 1), P, L. *Integer Flows and Cycle Covers of Graphs.* Cun-Quan Zhang. Pure & Appl. Math., V. 205. Marcel Dekker, 1997, xii + 379 pp, $145. [ISBN 0-8247-9790-6] The problem of face-coloring a planar graph can be formulated in terms of integer flows or cycle covers. This introduction presents main graph-theoretic results and open

problems. Requires only basic background in graph theory. LB

Number Theory, T(13–14). *The Theory of Remainders.* Andrea Rothbart. Janson, 1995, xii + 178 pp, $19.95 (P). [ISBN 0-939765-82-9] Whimsical, but substantial, look at modular structures narrated by Ant (Amateur Number Theorist) and Gnam (Game Nut and Magician). Developed to give middle school teachers a conceptual and connected look at authentic algebra. Suitable for liberal arts mathematics students as well as sophomore majors in "transition" courses. Exercises advance conceptual development. Optional problems accommodate different levels of sophistication. MW

Linear Algebra, T(18: 2), P. *Matrix Analysis.* Rajendra Bhatia. Grad. Texts in Math., V. 169. Springer-Verlag, 1997, xi + 347 pp, $49.95. [ISBN 0-387-94846-5] Advanced matrix theory with an emphasis on functional analytic properties. An excellent resource for matrix inequalities. BK

Algebra, T(16–17: 1), S, L.** *Geometry of the Quintic.* Jerry Shurman. Wiley, 1997, xi + 200 pp, $39.95 (P). [ISBN 0-471-13017-6] Interesting exploration of a single problem; draws on a range of (usually disparate) undergraduate topics: Riemann sphere and analysis, groups and algebra, icosahedral (and basic algebraic) geometry; uses ideas from Felix Klein through recent work on iterative solution methods. A nice "capstone" read for an ambitious math major. RM

Algebra, P. *Finite Fields: Normal Bases and Completely Free Elements.* Dirk Hachenberger. Intern. Ser. Eng. & Comp. Sci. Kluwer Academic, 1997, xii + 171 pp, $87.50. [ISBN 0-7923-9851-3] Considers the characterization, enumeration, and explicit construction of completely free elements in arbitrary finite dimensional extensions over finite fields. Assumes familiarity with basic Galois theory and finite field theory. CEC

Algebra, P. *A Survey of Modern Algebra.* Garrett Birkhoff, Saunders MacLane. AK Peters, 1997, xi + 500 pp, $59. [ISBN 1-56881-068-7] A corrected reprint of the *Fourth Edition* (TR, August–September 1977) by a new publisher.

Algebra, T(15: 1), S, L. *Numbers and Symmetry: An Introduction to Algebra.* Bernard L. Johnston, Fred Richman. CRC Pr, 1997, 260 pp, $29.95 (P). [ISBN 0-8493-0301-X] Introduces mathematical structures in concrete settings. Gaussian integers and polynomials are emphasized in the study of rings. Symmetry, to the extent of classifying wallpaper patterns, is the motivator for groups; linear algebra is used to study error-correcting codes over finite fields. CEC

Complex Analysis, T(17: 1). *Topics in Complex Analysis.* Mats Andersson. Universitext. Springer-Verlag, 1997, viii + 157 pp, $32.50 (P). [ISBN 0-387-94754-X] Moves quickly through fundamental results and focuses on connections to real analysis and harmonic analysis, e.g., Fatou's theorem, Nevanlinna theory, corona theory, H^p theory, and H^1-BMO duality. Assumes background in real analysis, integration theory, and functional analysis. PG

Complex Analysis, T*(16–17: 2). *Complex Analysis, Second Edition.* Joseph Bak, Donald J. Newman. Undergrad. Texts in Math. Springer-Verlag, 1997, x + 294 pp, $39.95. [ISBN 0-387-94756-6] No major changes in this edition. Still a very attractive textbook. (*First Edition*, TR, April 1983; Extended Review, March 1990.) TAV

Differential Equations, T*(16–17), L. *Nonlinear Differential Equations and Dynamical Systems, Second Revised and Expanded Edition.* Ferdinand Verhulst. Universitext. Springer-Verlag, 1996, x + 303 pp, $32.50 (P). [ISBN 0-540-60934-2] A nonponderous text for a second course in ODEs. Covers basic concepts of nonlinear DEs and stability analysis; defers bifurcation theory, chaos, and Hamiltonian systems to the end. Many examples and exercises with answers and hints. Few applications, but a nice theoretical text for those mainly interested in applications. (*First Edition*, TR, November 1990.) DK

Differential Equations, S(14), C. *The Maple O.D.E. Lab Book.* Darren Redfern, Edgar Chandler. Springer-Verlag, 1996, ix + 160 pp, $29.95 (P), with disk. [ISBN 0-387-94733-7] Introduction to the use of Maple V Release 4 for solving and analyzing first-order ODEs and systems. Includes explicit and implicit solutions, direction fields, phase planes, matrix methods, numerical methods, and applications. PG

Differential Equations, T(14). *Differential Equations with Boundary-Value Problems, Fourth Edition.* Dennis G. Zill, Michael R. Cullen. Brooks/Cole, xv + 652 pp, $76. [ISBN 0-534-95580-0] Changes from *Third Edition* include increased emphasis on using technology, non-linear equations and systems, boundary value problems, and modeling. (*Second Edition*, February 1989.) SK

Differential Equations, P. *A Treatise on Differential Equations, Sixth Edi-*

tion. A.R. Forsyth. Dover, 1996, xviii + 583 pp, $14.95 (P). [ISBN 0-486-69314-7] Unabridged republication of the 1956 printing of the 1929 Macmillan & Co. edition.

Differential Equations, P. *Catastrophe Theory and Its Applications.* Tim Poston, Ian Stewart. Dover, 1996, xviii + 491 pp, $18.95 (P). [ISBN 0-486-69271-X] Unabridged republication of the 1956 printing of the 1929 Macmillan & Co. edition. (1978 Pitman hardcover edition, TR, August–September 1978.)

Dynamical Systems, T(14–16), L.** *Chaos: An Introduction to Dynamical Systems.* Kathleen T. Alligood, Tim D. Sauer, James A. Yorke. Springer-Verlag, 1997, xvii + 603 pp, $39 (P). [ISBN 0-387-94677-2] An excellent text for an introductory dynamics course. Informal, but precise, mathematical style; quite accessible and motivating. Wonderful chapter-by-chapter "Lab Visits" (diverse applications of the concepts) and "Challenges" (short-term projects). Highly recommended. DK

Dynamical Systems, T(16–18), L. *Introduction to the Theory of Stability.* David R. Merkin. Transl. & Ed.: Fred F. Afagh, Andrei L. Smirnov. Texts in Appl. Math., V. 24. Springer-Verlag, 1997, xx + 319 pp, $49.95. [ISBN 0-387-94761-2] Translation of the *Third Edition* of a Russian text. Heavy emphasis on applications; roughly 1/4 of the text is given over to engineering and science applications and examples. DK

Numerical Analysis, S(16–18), P. *Dynamical Systems and Numerical Analysis.* A.M. Stuart, A.R. Humphries. Cambridge Univ Pr, 1996, xxii + 685 pp, $59.95. [ISBN 0-521-49672-1] An in-depth research monograph suitable for graduate courses in numerical analysis. Treats numerical computation as a dynamical system and examines two main issues: convergence of numerical attracting sets to true sets; preservation of structural properties (e.g., the Hamiltonian) by the numerical method. Examples throughout; many exercises. DK

Numerical Analysis, P. *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations.* Barry F. Smith, Petter E. Bjørstad, William D. Gropp. Cambridge Univ Pr, 1996, xii + 224 pp, $39.95. [ISBN 0-521-49589-X] Domain decomposition algorithms are used to break a PDE into subproblems to facilitate parallel computation of the solution. This book presents many domain decomposition algorithms and discusses their implementation. JO

Operator Theory, P. *Haar Series and Linear Operators.* Igor Novikov, Evgenij Semenov. Math. & Its Applic., V. 367. Kluwer Academic, 1997, xv + 218 pp, $118. [ISBN 0-7923-4006-X] Treats unconditional convergence; Fourier-Haar coefficients; reproducibility; martingales; monotone bases; criterion of equivalence of Haar and Franklin systems. Sadly, a quick reading caught several errors—a typo in the preface, a misspelling in the acknowledgement, an incorrect reference to an item in the bibliography (which, incidentally, contains 354 items!). KS

Functional Analysis, P. *The Descriptive Set Theory of Polish Group Actions.* Howard Becker, Alexander S. Kechris. London Math. Soc. Lect. Note Ser., V. 232. Cambridge Univ Pr, 1996, xi + 136 pp, $34.95 (P). [ISBN 0-521-57605-9]

Analysis, T(14: 1, 2), L*.** *An Introduction to Analysis.* Gerald G. Bilodeau, Paul R. Thie. Intern. Ser. in Pure & Appl. Math. McGraw-Hill, 1997, xi + 225 pp, $68.75. [ISBN 0-07-005662-5] A concise introductory text for a beginning analysis course. The topics are standard, from the development of the reals through uniform convergence, but the writing is especially clear. Extensive problem sets with interesting and well-chosen problems. A very attractive option; well worth considering. TAV

Analysis, T(16–18), S, L. *A Mathematical Introduction to Wavelets.* P. Wojtaszczyk. London Math. Soc. Stud. Texts, V. 37. Cambridge Univ Pr, 1997, xii + 261 pp, $21,95 (P); $59.95. [ISBN 0-521-57894-9; 0-521-57020-4] Though the literature on the 'newish' subject of wavelets is enormous, there is still a lack of readable introductory texts. This is one of many new books aimed at filling this gap. It is intended for (pure) math students who have a solid background in real analysis, and a knowledge of Lebesgue theory, Fourier series, and Hilbert spaces. KS

Analysis, P. *A Course of Modern Analysis, Fourth Edition.* E.T. Whittaker, G.N. Watson. Cambridge Univ Pr, 1996, 608 pp, $49.95 (P). [ISBN 0-521-58807-3] A reprint of the 1927 edition.

Analysis, P. *Asymptotic Attainability.* A.G. Chentsov. Math. & Its Applic., V. 383. Kluwer Academic, 1997, xiv + 322 pp, $156. [ISBN 0-7923-4302-6]

Analysis, T?(18: 2). *Measures and Probabilities.* Michel Simonnet. Universitext. Springer-Verlag, 1996, xiii + 510 pp, $44 (P). [ISBN 0-387-94644-6] The author describes his approach to measure and integration as an "intro-

ductory, yet sophisticated treatment." It is introductory in the sense that no previous course in measure is assumed, but the rudiments of topology and functional analysis are. It certainly is sophisticated as it uses the development of Daniell measure to provide results on Radon as well as abstract measures on sets. In three parts: Daniell Measures, Measures on Subrings, and Convergence of Random Variables. TAV

Algebraic Geometry, T(18), P. *Lectures on Vector Bundles.* J. Le Potier. Transl: A. Maciocia. Stud. in Adv. Math., V. 54. Cambridge Univ Pr, 1997, viii + 251 pp, $59.95. [ISBN 0-521-48182-1] Lectures from two courses on moduli spaces of vector bundles. Part 1 concerns the classification of vector bundles on algebraic curves. Part 2 concerns semi-stable sheaves on the projective plane. Assumes background in algebraic geometry, Chern classes, and algebraic sheaves. JD

Differential Geometry, T(18: 1), P. *Conformal Differential Geometry and Its Generalizations.* Maks A. Akivis, Vladislav V. Goldberg. Pure & Appl. Math. Wiley, 1996, xiv + 383 pp, $69.96. [ISBN 0-471-14958-6] A conformal space is, roughly, an n-dimensional sphere equipped with the group of conformal transformations on the sphere. Conformal differential geometry concerns such spaces and related ideas. This book gives a thorough introduction to the subject. JO

Algebraic Topology, S(17), P, L. *A User's Guide to Algebraic Topology.* C.T.J. Dodson, Phillip E. Parker. Math. & Its Applic., V. 387. Kluwer Academic, 1997, xii + 405 pp, $209. [ISBN 0-7923-4292-5] Not a text—many proofs are merely referenced. Covers extension and lifting problems, homotopy, cohomology, sheaves, bundles, obstruction theory. Appendices on algebra and general topology. Applications to theoretical physics. JD

Topology, P. *The Hauptvermutung Book: A Collection of Papers of the Topology of Manifolds.* Ed: A.A. Ranicki, *et al.* K-Mono. in Math., V. 1. Kluwer Academic, 1996, vi + 190 pp, $108. [ISBN 0-7923-4174-0] The *Hauptvermutung* is the conjecture that any two triangulations of a polyhedron (or a manifold) are combinatorially equivalent. This book brings together the literature on the subject, including proofs of several results that have been assumed but unpublished up to now. JO

Topology, T(18), P. *Theory of Degrees with Applications to Bifurcations and Differential Equations.* Wieslaw Krawcewicz, Jianhong Wu. Canadian Math. Soc. Ser. of Mono. & Adv. Texts. Wiley, 1997, xiv + 374 pp, $89.95. [ISBN 0-471-15740-6] A unified treatment of Brouwer (and extensions), Dold-Ulrich, and S^1-degrees and their relation to the existence of bifurcations of ODEs. Requires some knowledge of analysis, topology, and ODEs. SK

Optimal Control, P. *Geometric Control Theory.* Velimir Jurdjevic. Stud. in Adv. Math., V. 51. Cambridge Univ Pr, 1997, xviii + 492 pp, $79.95. [ISBN 0-521-49502-4]

Optimal Control, P. *Nonlinear $\mathcal{H}_\infty$ Control: The Singular Case.* W.C.A. Maas. CWI Tract, No. 118. Stichting Mathematisch Centrum, 1996, v + 197 pp, Dfl. 40 (P). [ISBN 90-6196-468-7]

Probability, T*(17–18: 1, 2). *A Modern Approach to Probability Theory.* Bert Fristedt, Lawrence Gray. Prob. & Its Applic. Birkhäuser Boston, 1997, xx + 756 pp, $64.50. [ISBN 0-8176-3807-5] Another possible title is "Everything You Need to Know About the Theory of Probability." This ambitious work covers the essentials in a clear and readable fashion. From Σ-fields to sums and convergences to interacting particle systems with all the stops in between. A must for professionals and an attractive text for a graduate course. Extensive appendices, credits, and comments. TAV

Statistical Methods, T(17–18: 1). *Applied Factor Analysis in the Natural Sciences.* Richard A. Reyment, K.G. Jöreskog. Cambridge Univ Pr, 1996, xii + 371 pp, $39.95 (P); $54.95. [ISBN 0-521-57556-7; 0-521-41242-0] Provides an introduction to multivariate data analysis, aspects of linear algebra, and factor analysis. Discusses R-mode, Q-mode, and Q-R-mode methods. Contains examples and case histories from geology and ecology. Appendix contains MATLAB computer programs. RS

Statistical Methods, T(17–18), P. *Nonparametric Methods for Quantitative Analysis, Third Edition.* Jean Dickinson Gibbons. American Sciences Pr, 1997, xvi + 537 pp, $70.95 (P). [ISBN 0-935950-37-0] New edition incorporates computer package (MINITAB, SAS, SPSS) solutions in many examples. Several new topics; expanded problem sets. (*Second Edition*, TR, November 1986.) RS

Statistics, P. *Statistical Test Limits in Quality Control.* G.D. Otten. CWI Tract, No. 116. Stichting Mathematisch Centrum, 1996, vi + 144 pp, Dfl. 35 (P). [ISBN 90-6196-469-5]

Algorithms, T(18: 1), P. *Parallel Algorithms for Regular Architectures: Meshes and Pyramids.* Russ Miller, Quentin F. Stout. MIT Pr, 1996, xvii + 310 pp, $40. [ISBN 0-262-13233-

8] Good presentation of parallel algorithms for mesh and pyramid parallel architectures. Considers sorting and matrix multiplication as well as many problems from image processing, computational geometry, and graph theory. JO

Theory of Computation, T(15: 1), L. *Languages and Machines: An Introduction to the Theory of Computer Science, Second Edition.* Thomas A. Sudkamp. Addison-Wesley, 1997, xv + 569 pp, $45.95. [ISBN 0-201-82136-2] Good exposition, proofs, and coverage. Connections to practical computing include a section on a context-free grammar describing Pascal, and sections on $LL(k)$ and $LR(k)$ grammars. JO

Computer Science, P. *Logic Programming.* Ed: Michael Maher. MIT Pr, 1996, xix + 554 pp, $85 (P). [ISBN 0-262-63173-3] Proceedings of the 1996 Joint International Conference and Symposium on Logic Programming held in Bonn, Germany.

Computer Science, P. *Computational Differentiation: Techniques, Applications, and Tools.* Eds: Martin Berz, *et al.* SIAM, 1996, xv + 421 pp, $65 (P). [ISBN 0-89871-385-4] Papers from the Second International Workshop on Computational Differentiation held in February 1996 in Santa Fe, New Mexico.

Computer Science, T(15–17: 1), S, P.** *Algebra of Programming.* Richard Bird, Oege de Moor. Intern. Ser. in Comp. Sci. Prentice Hall, 1997, xiv + 295 pp. [ISBN 0-13-507245-X] One may calculate, algebraically, a program for a class of problems (parametrized by a data type) whose specification satisfies some structure by proving a theorem which asserts a strategy (greedy, dynamic programming, etc.) applies, then algebraically instantiating the algorithm (e.g., from the proof). Nice development based on categorical calculus of relations instantiated into functional programming language. RM

Computer Science, P. *Graph Reduction on Shared-Memory Multiprocessors.* K.G. Langendoen. CWI Tract, No. 117. Stichting Mathematisch Centrum, 1996, vii + 199 pp, Dfl. 40 (P). [ISBN 90-6196-470-9]

Applications (Biological Science), P. *Next Generation Environmental Models and Computational Methods.* Eds: George Delic, Mary F. Wheeler. SIAM, 1997, xii + 375 pp, $65 (P). [ISBN 0-89871-378-1] Proceedings of a 1995 workshop at the National Environmental Supercomputing Center, Bay City, Michigan. Topics include global and regional circulation models, aquatic systems, groundwater transport of contaminants, and inverse problem methods.

Applications (Communication Theory), P. *Chaotic, Fractal, and Nonlinear Signal Processing.* Ed: Richard A. Katz. American Institute of Physics Pr, 1996, xxi + 847 pp, $165. [ISBN 1-56396-443-0] Technical papers from the ONR/NUWC Third Technical Conference on Nonlinear Dynamics and Full-spectrum Processing held in 1995 in Mystic, Connecticut.

Applications (Mechanics), P. *Global Bifurcation in Variational Inequalities: Applications to Obstacle and Unilateral Problems.* Vy Khoi Le, Klaus Schmitt. Appl. Math. Sci., V. 123. Springer-Verlag, 1997, xiv + 250 pp, $59.95. [ISBN 0-387-94886-4]

Applications (Quantum Theory), P. *Proceedings Seminar 1989–1990: Mathematical Structures in Field Theory.* Eds: E.A. de Kerf, H.G.J. Pijls. CWI Syllabus, No. 39. Stichting Mathematisch Centrum, 1996, 163 pp, Dfl. 40 (P). [ISBN 90-6196-448-2] Selected lectures from a seminar at the University of Amsterdam.

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–223: Experimental Robotics IV.* Eds: O. Khatib, J.K. Salisbury. Springer-Verlag, 1997, xix + 574 pp, $86 (P). [ISBN 3-540-76133-0] Proceedings of the Fourth International Symposium held at Stanford University in 1995.

Applications, P. *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements.* S. Twomey. Dover, 1996, x + 243 pp, $9.95 (P). [ISBN 0-486-69451-8] Republication, with corrections, of Volume 3 of *Developments in Geomathematics* published by Elsevier Scientific in 1977.

Applications, P. *Chaos and the Changing Nature of Science and Medicine: An Introduction.* Eds: Donald E. Herbert, *et al.* American Institute of Physics Pr, 1996, xi + 203 pp, $130. [ISBN 1-56396-442-2] Papers from a 1995 workshop in Mobile, Alabama.

Reviewers

LB: Lynne Baur, Carleton; LC: Laura Chihara, St. Olaf; CEC: Clifton E. Corzatt, St. Olaf; JD: Jill Dietz, St. Olaf; PG: Philip Gloor, St. Olaf; DK: Danny Kaplan, Macalester; SK: Steve Kennedy, Carleton; BK: Brenda Kroschel, Macalester; RM: Richard Molnar, Macalester; JO: Jeff Ondich, Carleton; AO: Arnold Ostebee, St. Olaf; KS: Karen Saxe, Macalester; RS: Richard Single, St. Olaf; KES: Kay E. Smith, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf; TAV: Theodore A. Vessey, St. Olaf; MW: Martha Wallace, St. Olaf.

THE AUTHORS

NORBERT PEYERIMHOFF received his Ph.D. in 1993 at the University of Augsburg, under the direction of Jochen Brüning. In 1994–1996, as a postdoctorate fellow by the Deutsche Forschungsgemeinschaft (DFG), he enjoyed the great hospitality of the Graduate Center of the City University of New York (CUNY) and, in particular, the differential geometry group. In New York, he enjoyed lessons in ballroom and Latin dancing and spent a lot of money on mathematics books. Recently, he accepted a position as Assistant at the University of Basel.

DON ZAGIER is half American and half European: he was born both in the United States and in Germany (yes, there is a way), graduated from secondary school in both California and England, received his university education in both continents (B.Sc.'s in physics and in mathematics from MIT and a D.Phil. granted by Oxford University but supervised by F. Hirzebruch in Bonn), and for many years held simultaneous jobs in Bonn and Maryland. At present the European half is in the ascendant and he is attached to the University of Utrecht in Holland (as a permanent, but part-time, professor) and to the Max Planck Institute for Mathematics in Bonn (where he is now doing time as Managing Director). His primary field of interest is number theory, but one can get him to work in any domain.

DAVID CALLAN was born in Ireland and did his undergraduate work at University College, Dublin and graduate work at the University of Notre Dame. He now lives in Madison, Wisconsin (rated #1 by Money Magazine) and maintains interests in problem solving, matrix theory, combinatorics, and actuarial science. He enjoys travel, near and far, especially by bicycle.

BETTINA RICHMOND received her Vordiplom from Universität Würzburg (Germany). She finished her Ph.D. at Florida State University in 1985 under the direction of Warren D. Nichols. Since 1986 she has taught at Western Kentucky University, where she is professor of mathematics. Her primary area of research is structure theory of Hopf algebras.

TOM RICHMOND received his Ph.D. from Washington State University in 1986 under the direction of Darrell C. Kent. Since that time he has been at Western Kentucky University where he is professor of mathematics. His research interests include partial orders and topology.

ALBERT FATHI obtained his Ph.D. from Université Paris-Sud in 1980. He held positions in CNRS-Université Paris-Sud and University of Florida, Gainesville. He is now at Ecole Normale Supérieure de Lyon. His main interests are dynamical systems and geometry.

JEFF KNISLEY received his Ph.D. at Vanderbilt University in 1990, and is now at East Tennessee State University in Johnson City, Tennessee. His interests include applied mathematics and operator theory, and he is working with a neurophysiologist to study mathematical models of the flow of electricity through the dendrites of a neuron.

ARNOLD OSTEBEE was raised on a farm in Iowa. As an undergraduate at St. Olaf College, he majored in mathematics and physics. Subsequently, he completed his Ph.D. in mathematical physics (rigorous statistical mechanics) at SUNY/Stony Brook. He returned to St. Olaf as a member of the faculty in 1980. He now serves as chair of St. Olaf's mathematics department and as the editor of the Telegraphic Reviews section of the MONTHLY.

PAUL ZORN was born in India, and had his primary and secondary schooling there. He did his undergraduate work at Washington University in St. Louis, and his Ph.D., in complex analysis, at the University of Washington, Seattle. Since 1981 he has been on the mathematics faculty at St. Olaf College. His editorial work includes service with FOCUS, the New Mathematical Library, and the 1994 and 1995 editions of *What's Happening in the Mathematical Sciences*. He is the editor of *Mathematics Magazine*.

JIM KAPUT began professional life as a mathematician working in category theory with a Ph.D. from Clark University under John Kennison in 1968. In the early 1970s he turned his attention to college student mathematical learning and thinking. Since then he has been tracing the developmental and semiotic path of mathematical learning backwards to earlier grade levels, and examining the representational roles of technology in learning, teaching, and thinking. He also enjoys wearing out running shoes and hiking boots.

RICHARD ASKEY started to learn calculus out of two books, the then mainstream book by Sherwood and Taylor and a very old fashioned book that had been used at West Point early in the century, to judge by the comments in the margin. A bit of each stuck, but no longer taught techniques such as integration by undetermined coefficients really made it easier to understand such things as the fundamental theorem of calculus. As a result, he has always wondered about the wisdom of trying to teach ideas without teaching technical skills at the same time.

JOSEPH KUPKA was born in 1942 and is not dead yet. He received a Ph.D. in analysis from UC Berkeley in 1970 and has been teaching at Monash ever since. He expects that the Berkeley math professors would be aghast were they to learn of his post-burnout interests in combinatorial problems of elementary character, not to mention his continuing campaign to demonstrate that J. S. Bach really did have a sense of humor.

APOLONIUSZ TYSZKA was graduated from Jagellonian University.

DAN CHRISTENSEN received his bachelor's degree in pure and applied mathematics from the University of Waterloo after toying with computer science for a few years. He is completing graduate studies at M.I.T. in homotopy theory, and occasionally finds time to hike the nearby hills and paddle the rivers. The work described in his article was done during the summer of 1995 when Dan was supervising Mark in the Research Science Institute program at MIT.

MARK TILFORD is a member of the class of 2000 at Caltech. He completed calculus in ninth grade and continued with math classes at Maryville University and Washington University. During the summers, he studied at TIP at Duke, YSP at Rose-Hulman, and the Research Science Institute, sponsored by the Center for Excellence in Education and MIT, where he worked with Dan Christensen on the subset take-away problem. In 1996, Mark received an Honorable Mention on the USAMO and Honorable Mention on the Putnam Competition. In his spare time, he collects mysteries, and books and articles by Martin Gardner.

RONALD MICKENS received the Ph.D. in theoretical physics from Vanderbilt University in 1968. For the next two years he held a National Science Foundation Post-doctoral Fellowship at the Center for Theoretical Physics, Massachusetts Institute of Technology. After a decade of teaching and research at Fisk University, he assumed his present position as Callaway Professor of Physics at Clark Atlanta University. His main research interests are determining asymptotic properties of the solutions to difference equations, discrete modeling of continuous systems, nonlinear oscillations, and the history of science.

WAYNE ROBERTS has been a member of the MAA committee, Calculus Reform and the First Two Years (CRAFTY) over the entire decade of calculus reform, and was Chair of that committee from 1992 to 1995. He is editor of the new CRAFTY book, *Calculus: The Dynamics of Change* and he directed the project that developed the five-volume *Resources for Calculus* published by the MAA. A member of the Macalester College faculty for thirty years, and for eight years Chair of the Department of Mathematics and Computer Science, Roberts was appointed Provost of the College in July 1995. He was awarded the MAA North Central Section Citation for Meritorious Service in 1996.

EDITOR'S ENDNOTES

This month's four-article mini-forum on calculus reform addresses important issues of concern to everyone involved in collegiate mathematics. We are grateful to Jeff Knisley, Paul Zorn, Arnold Ostebee, James Kaput, and Richard Askey for their efforts.

Two recent paperback books contain a wealth of sensible advice for mathematical authors:

Handbook of Writing for the Mathematical Sciences, by Nicholas J. Higham. Society for Industrial and Applied Mathematics, Philadelphia, 1993, 241 pp., \$21.50. ISBN 0-89871-314-5.

A Primer of Mathematical Writing, by Steven G. Krantz. American Mathematical Society, Providence, 1996, 223 pp., \$19.00. ISBN 0-8218-0635-1.

Members may obtain these books from their respective societies at a discount.

Readers who have enjoyed filler items containing MONTHLY extracts from 100, 50, and 25 years ago have Joseph Gallian to thank for their pleasure. Joe reports that he selected "short things that were amusing, quaint, or curious. Included are problems, book reviews, or simply titles of books reviewed, and announcements of various sorts. In the case of the problems, some were chosen because of whom they were submitted by (e.g., Pólya, Dickson) while others were chosen to illustrate the contrast between the problems in current issues and those from 100 years ago."

Paolo Ribenboim reports the following comments about his article in the August, 1996 issue [103 (1996) 529–538] contributed by S. W. Golomb:

The quantitative version of the twin primes conjecture is not quite as stated on p. 531. The conjectured asymptotic form of the number of primes $p \leq N$ such that $p + 2$ is also prime is $C_2 N/(\log N)^2$, where the *twin prime constant* $C_2 = 1.32032\dots$ is a certain infinite product over odd primes [MONTHLY 67 (1960) 767–769].

According to L. E. Dickson's *History of the Theory of Numbers*, de Polignac's conjecture is not merely "that every even number is the difference of two primes" (p. 531), but that every even number is a difference of two primes in infinitely many ways. Thus, de Polignac's conjecture includes the twin primes conjecture as a special case.

Our December, 1996 issue contained an unattributed filler item (see p. 916) extracted from the winter, 1991 issue of *The Mathematical Intelligencer*. George F. Simmons points out that original source of this item is the preface to his book *Differential Equations with Applications and Historical Notes*, second edition, McGraw-Hill, Inc., New York, 1991. The full quotation (see p. xix) is:

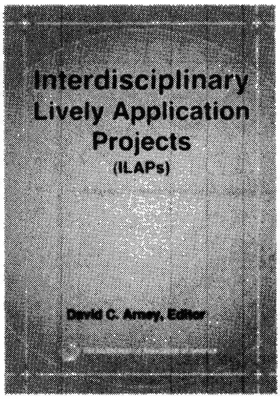
Some will think that a mathematical argument either is a proof or is not a proof. In the context of elementary analysis I disagree, and believe instead that the proper role of a proof is to carry reasonable conviction to one's intended audience. It seems to me that mathematical rigor is like clothing: in its style it ought to suit the occasion, and it diminishes comfort and restricts freedom of movement if it is either too loose or too tight.

Leonard Gillman writes the following about Rethinking Rigor in Calculus in the March issue [104 (1997) 231–240]:

Tom Tucker recommends avoiding *deus-ex-machina* auxiliary functions in proofs, such as a proof of the Mean Value Theorem. Of course. Let f and g be continuous on $[a, b]$ and differentiable on (a, b), and suppose f and g agree at a and b. Then $f - g$ satisfies the hypotheses of Rolle's Theorem, so there is a point in (a, b) where f' and g' agree. The Mean Value Theorem is the special case in which the graph of g is a straight line. This argument is in the calculus text by Ford and Ford, McGraw-Hill, 1963.

Roger A. Horn, *Editor*

THE MATHEMATICAL ASSOCIATION OF AMERICA



Interdisciplinary Lively Application Projects (ILAPs)

Series: Classroom Resource Materials

David C. Arney

Interdisciplinary Lively Application Projects (ILAPs) are small group projects developed through the cooperation of faculty from mathematics and partner disciplines. These projects will provide teachers with material that can help their students understand mathematical concepts, develop strong mathematical skills and motivate them towards an interest in future subjects accessible through the study of mathematics. It is an important step towards helping students acquire a broad, interdisciplinary outlook towards problem solving.

The ILAPs provide supplemental classroom resource materials in the form of eight project handouts that you can use as student homework assignments. They require students to use scientific and quantitative reasoning, mathematical modeling, symbolic manipulation skills, and computational tools to solve and analyze scenarios, issues, and questions involving one or more disciplines. Sample solutions to the problems, background material, notes to the instructor and a student writing guide are also included.

The prerequisite skills for the eight projects presented in the book range from freshmen-level algebra, trigonometry, and precalculus; through calculus, elementary and intermediate differential equations, and discrete mathematics to advanced calculus and partial differential equations. The partner disciplines included in the projects are: mechanics, physics, chemistry, engineering, geography, topography, and exercise physiology. You can use the projects as a supplement to a textbook in any of the following undergraduate areas: precalculus, calculus, linear algebra, differential equations, discrete mathematics, mathematical modeling, advanced calculus, partial differential equations, and numerical computing.

The book also contains several supporting articles that describe uses for these projects.

Contents: **ILAPs:** Getting Fit with Mathematics; Decked Out; Parachute Panic; Flying with Differential Equations; Planning a Backpacking Trip to Pikes Peak; SMOG in Los Angeles Basin, Structural Mechanics — Beams and Bridges; Contaminant Transport. **Articles:** Technical Report Format and Writing Guide; Project INTERMATH: An Interdisciplinary Approach to Cultural Change; ILAP Products: Authoring, Testing and Editing; Interdisciplinary and Integrated Curriculum Models; Interdisciplinary Communication and Understanding; Interdisciplinary Projects at West Point.

Catalog Code: ILAP/JR
206 pp., Paperbound, 1997
ISBN 0-88385-706-5
List: $27.50 MAA Member: $ 22.00

Phone in Your Order Now!* ☎ *1-800-331-1622

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders:** $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ______________________

Address ______________________

City ____________ State _____ Zip ________

Phone ______________________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	ILAP/JR	______	______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

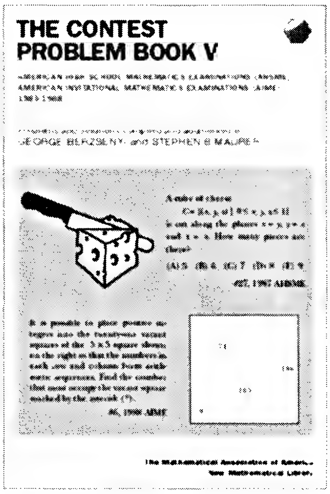
Shipping & handling ______

TOTAL ______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________



The Contest Problem Book V

American High School Mathematics Examinations and American Invitational Mathematics Examinations, 1983–1988

Series: New Mathematical Library

George Berzsenyi and Stephen B Maurer

Over the years perhaps the most popular of the MAA problem books have been the high school contest books, covering the yearly American High School Mathematics Examinations (AHSME) that began in 1950, co-sponsored from the start by the MAA. Book V also includes the first six years of the American Invitational Mathematics Examination (AIME) which was developed as an intermediate step between the AHSME and the USA Mathematical Olympiad (USAMO). The AIME has a unique answer format — all answers are integers between 0 and 999.

The editors of this volume, George Berzsenyi and Stephen B Maurer, were respectively the chair of the AIME and the AHSME during this period. In addition to a thorough index, they have added much material not included in Contest Books I–IV:

- a comprehensive guide to other problem materials world wide,
- additional solutions,
- dropped problems,
- statistical information,
- information on test development and history.

This volume is a must for avid fans of elementary problems.

Contest Books I–IV appear as NML volumes 5, 17, 25, and 29.

Catalog Code: NML-38/JR
308 pp., Paperbound, 1997
ISBN 0-88385-640-9
List: $24.95 MAA Member: $20.00

Phone in Your Order Now! ☎ *1-800-331-1622*

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders**: $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ____________________

Address ____________________

City ____________ State _____ Zip ________

Phone ____________________

QTY.	CATALOG CODE	PRICE	AMOUNT
_______	NML-38/JR	_______	_______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling ________

TOTAL ________

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________________ Expires ___/___

Signature ____________________

THE AMERICAN MATHEMATICAL
MONTHLY

Volume 104, Number 9 November 1997

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and Notes should be sent to the Editor:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send your email address and 3 copies of the complete manuscript (including all figures with captions and lettering), typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper.

Letters to the Editor on any topic are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

See the MONTHLY section of MAA Online for current information such as contents of issues, descriptive summaries of forthcoming articles, tips for authors, and preparation of manuscripts in TEX:

http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all problems/solutions material, typewritten on only one side of the paper.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Mr. JOE WATSON, Advertising Manager

Change of address, missing issues inquiries, and other subscription correspondence:
MAA Service Center
maahq@maa.org

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Mathematics, Statistics, and Teaching

George W. Cobb and David S. Moore

How does statistical thinking differ from mathematical thinking? What is the role of mathematics in statistics? If you purge statistics of its mathematical content, what intellectual substance remains?

In what follows, we offer some answers to these questions and relate them to a sequence of examples that provide an overview of current statistical practice. Along the way, and especially toward the end, we point to some implications for the teaching of statistics.

1. INTRODUCTION: AN OVERVIEW OF STATISTICAL THINKING. Statistics is a methodological discipline. It exists not for itself but rather to offer to other fields of study a coherent set of ideas and tools for dealing with data. The need for such a discipline arises from *the omnipresence of variability.* Individuals vary. Repeated measurements on the same individual vary. In some circumstances, we want to find unusual individuals in an overwhelming mass of data. In others, the focus is on the variation of measurements. In yet others, we want to detect systematic effects against the background noise of individual variation. Statistics provides means for dealing with data that take into account the omnipresence of variability.

1.1. The role of context. The focus on variability naturally gives statistics a particular content that sets it apart from mathematics itself and from other mathematical sciences, but there is more than just content that distinguishes statistical thinking from mathematics. Statistics requires a different *kind* of thinking, because *data are not just numbers, they are numbers with a context.*

Example 1. The mystery of Andover. The finite sequence (3, 5, 23, 37, 6, 8, 20, 22, 1, 3) shows a distinctive pattern when plotted (Figure 1) but the numbers and the pattern have no meaning or interest until we know their context. They are in fact monthly totals of people formally accused of witchcraft in Essex County, Massachusetts, beginning in February, 1692. The plot shows two waves of accusations, separated by a low point in the summer of 1692. The pattern becomes still more meaningful when we know that the first hanging of a convicted witch (Bridget Bishop) took place June 10, 1692: it is not hard to imagine the sobering effect of that first execution in the small community of Salem Village (now Danvers). But why the second wave of accusations? It turns out that the accusations in the first wave were directed against residents of Salem Village, Salem Town, and all but one of the half-dozen immediately adjacent towns; in the second wave the majority of the accusations were directed against residents of the one other adjacent town, Andover. Our sources [**3**, **4**] do not provide much explanation for what happened in Andover, but the pattern, together with what we know of the context, tells at least part of a story and raises some interesting questions.

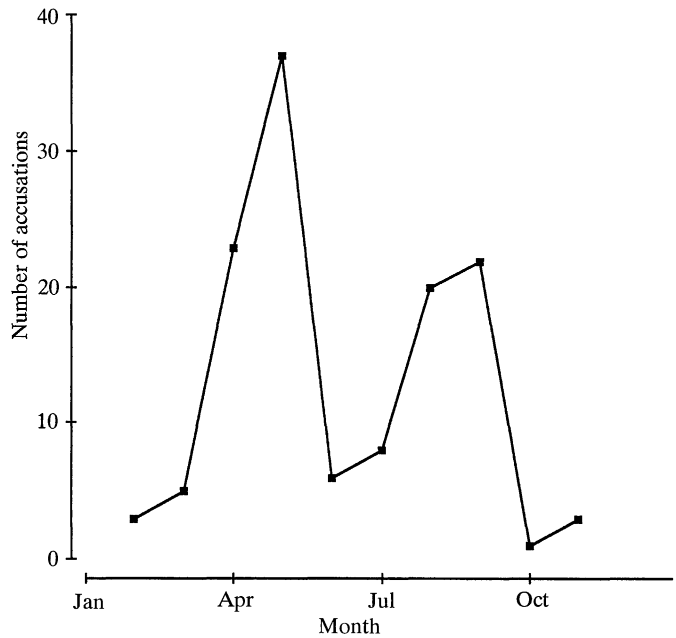


Figure 1. Numbers of people accused of witchcraft in Essex County, MA, 1692.

Although this first example has almost no mathematical content, its interplay between pattern and context is typical of the interpretive part of statistical thinking. For a more familiar example of a very different sort, consider testing that two normal distributions have equal means.

Example 2a. A model for comparing normal means. Consider the standard model involving two sets of independent, identically distributed (iid) random variables:

$$X_1, X_2, \ldots, X_n \text{ iid } N(\mu_1, \sigma_1) \qquad Y_1, Y_2, \ldots, Y_m \text{ iid } N(\mu_2, \sigma_2)$$

It follows that $\bar{x} = (\Sigma x_i)/n$ and $s_1^2 = \Sigma(x_i - \bar{x})^2/(n-1)$ are sufficient statistics for μ_1 and σ_1^2, with parallel results for the Ys. Informally, a statistic is *sufficient* for a parameter if it uses all the information about that parameter contained in the sample. More formally, the conditional distribution of the data, given the sufficient statistic, doesn't depend on the parameter. The Rao-Blackwell Theorem guarantees that no unbiased estimator can have a smaller variance than one based on a sufficient statistic. Both $\bar{x}$ and s_1^2 are unbiased: $E(\bar{x}) = \mu_1$ and $E(s_1^2) = \sigma_1^2$. Finally, their joint distribution is known: the sample mean $\bar{x}$ is normal with variance σ_1^2/n, and, independently, $(n-1)s_1^2/\sigma_1^2$ is chi-square on $(n-1)$ degrees of freedom. Suppose now we want to test H_0: $\mu_1 = \mu_2$. If $\sigma_1^2 = \sigma_2^2$ then a sufficient and unbiased estimator for the common variance is obtained by pooling:

$$s_p^2 = \left[(n-1)s_1^2 + (m-1)s_2^2\right]/(n+m-2)$$

If H_0 is true, then $(\bar{x} - \bar{y})/s\sqrt{(1/n) + (1/m)}$ has a Student's t-distribution on $n + m - 2$ degrees of freedom, and we can use the value of t computed from the data to test the null hypothesis. If t is far enough from 0, we conclude that $\mu_1 \neq \mu_2$.

This example differs most strikingly from the first in two ways: mathematical content and the role of context. Example 1, which has essentially no mathematical content, finds its intellectual substance almost entirely in the interplay between pattern and story. Example 2, which has essentially no content apart from mathematics, gets it intellectual substance without any explicit reference to applied context.

Although mathematicians often rely on applied context both for motivation and as a source of problems for research, the ultimate focus in mathematical thinking is on abstract patterns: the context is part of the irrelevant detail that must be boiled off over the flame of abstraction in order to reveal the previously hidden crystal of pure structure. *In mathematics, context obscures structure.* Like mathematicians, data analysts also look for patterns, but ultimately, in data analysis, whether the patterns have meaning, and whether they have any value, depends on how the threads of those patterns interweave with the complementary threads of the story line. *In data analysis, context provides meaning.*

The difference has profound implications for teaching. To teach statistics well, it is not enough to understand the mathematical theory; it is not even enough to understand also the additional, non-mathematical theory of statistics. One must, like a teacher of literature, have a ready supply of real illustrations, and know how to use them to involve students in the development of their critical judgment. In mathematics, where applied context is so much less important, improvised examples often work well, and teachers of mathematics become skillful at inventing examples on the spot. (Need a function to illustrate the chain rule? No problem: just make one up.) In statistics, however, improvised examples don't work, because they don't provide authentic interplay between pattern and context. Much as Bertrand Russell likened mathematics to sculpture for the austerity of its abstraction, one might think of data analysis as like poetry, where pattern and context are inseparable. Imagine yourself teaching a lesson on basic prosody, introducing dactylic hexameter. It is not enough to say "TA ta ta, TA ta ta, TA ta ta, . . . ;" your students need to hear dactyls in a real poem **[20]**: "This is the forest primeval. The murmuring pines and the hemlocks." In a similar spirit, the teacher of statistics needs to know the data literature. If, for example, when you teach plots for data distributions, you use data on inter-eruption times for Old Faithful **[30]** and lengths of reigns of English kings and queens **[13]**, your students can learn more than just the methods themselves. The bimodal shape of the inter-eruption times suggests two kinds of eruptions, and the distribution of monarchs' reigns shows the skewness toward high values that is typical of waiting times.

The contrasting roles of context in mathematics and statistics, especially as illustrated in the deliberately extreme first two examples, might seem to lend support to the false implication in Bullock's **[5]** assertion that "Many statisticians now claim that their subject is something quite apart from mathematics, so that statistics courses do not require any preparation in mathematics." In fact, while we find the evidence that statistics is not mathematics persuasive (see **[22]**, **[24]**), all statistics courses require some preparation in mathematics, and some require a great deal. Elaborate mathematical theories undergird some parts of statistics, and the study of those theories is part of the training of statisticians. But although statistics cannot prosper without mathematics, the converse fails. That statistics is not a necessary part of a mathematician's training is implicit in the statement by the eminent probabilist David Aldous **[1]** that he "is interested in the applications of probability to all scientific fields *except statistics*."

What then, is the role of mathematics in the science of statistics? An answer should begin with a more systematic look at the logic of analyzing data.

1.2. A schematic overview of statistical analysis. An old-style course that wanted to be conscientious about applications might finish off the second example with a little coda of an exercise. The data, although not this invented exercise, are from [**25**]; the full study is described in [**21**].

Example 2b. Calcium and blood pressure. Does increasing the amount of calcium in our diet reduce blood pressure? The following numbers give the decrease after 12 weeks in systolic blood pressure for 21 human subjects. The 10 subjects in Group 1 took a calcium supplement for 12 weeks; the 11 in Group 2 took a placebo. Test the hypothesis that the calcium had no effect on blood pressure.

Group 1 (calcium):	$7, -4, 18, 17, -3, -5, 1, 10, 11, -2$
Group 2 (placebo):	$-1, 12, -1, -3, 3, -5, 5, 2, -11, -1, -3$

This exercise, put so tersely, is a caricature, one that encourages the mistaken view that once the mathematical derivations from a model are completed, applications are largely a matter of routine arithmetic. For a more realistic perspective, consider Figure 2, a diagram of the stages in a statistical analysis. Before considering this crude outline in detail, two cautions are essential.

```
       (1)      (2)
Design --> Data --> Patterns
                    Model(s) --> Methods --> Results --> Intrepretation
                            (3)         (4)         (5)
```

Figure 2. A schematic representation of the phases of data production and analysis.

1. The summary oversimplifies by suggesting a strict left-to-right progression. In reality, the process of data analysis is neither linear not unidirectional. Several transitions involve a dialog of sorts, sometimes between adjacent elements, but sometimes among more than just two. Thus, for example, the choice of design for data production determines the structure of the resulting data, but knowledge based on data already in hand can help shape the design, as when knowing the size of variation from one subject to another helps decide how many subjects will be needed. Similarly, the data may suggest a model, but the model leads to methods that send us back to the data to check for possible violations of the model's assumptions. Perhaps most important of all, as we shall see, the final stage, interpretation of the results, depends in a crucial way on the first stage, the kind of design used for producing the data.
2. The rough and qualified ordering of stages here is not meant to suggest that we think the topics taught in an introductory statistics course should follow the same order. For reasons presented later, we recommend beginning with methods for exploring and describing data, then going "back" to data production, and from there to formal inference.

With these cautions assumed, the flowchart can provide a useful framework for examining the role of mathematics in statistics and summarizing elements of the

non-mathematical substance of the subject. Here are four quick observations:

1. Design, exploration, and interpretation are core elements of statistical thinking. All three elements are heavily dependent on context, but at the *introductory* level they involve very little mathematics. The (largely non-mathematical) theory of experimental design is decades old and well developed; the theory of exploration is newer, and at present still primitive, although computer-based tools for exploration have become quite sophisticated; the theory of interpretation is fragmentary at best.
2. The classical course in mathematical statistics corresponds so neatly to transition (3) that "from models to methods" might almost serve as a course title. Context is largely irrelevant here, because models are presented abstractly, as in Example 2a, and a typical derivation simply applies one optimality principle or another (least squares, maximum likelihood) to deduce the method *de jour*.
3. Transition (4), from methods to results, is the focus of the old-style cookbook course, in which each method is summarized by a set of formulas. Context is irrelevant here also, in that you can learn computational algorithms, and in fact learn them more efficiently, if you resist any temptation to encumber your brain with concern about what the methods are good for. All the same, some courses have tried to make the throat-clogging bolus of rote easier to get down by sugar-coating it with a thin glaze of ersatz context. Fortunately, the computer is fast sweeping courses like these into the dustbin of curricular history.
4. It is perhaps ironic that transitions (3) and (4), the two that have most often been the focus of courses at the introductory level, are precisely the two that are intellectually most automatic (given our current limited understanding and less developed theory of the other transitions) and so offer the least room for judgment and creativity.

To develop these points in more detail, we return to the example of calcium and blood pressure. In what follows, we combine the stages of Figure 2 under three broader headings: data production, data analysis, and inference.

2. THE CONTENT OF STATISTICS

2.1. Data production. The standard model of Example 2a is incomplete in a most serious way: it does not distinguish between observational data (e.g., from a sample survey) and data from a randomized comparative experiment. This distinction, between observation and experiment, is one of the most important in statistics. Researchers often want to reach *causal* conclusions: calcium *causes* a reduction in blood pressure. Experiments often allow causal conclusions, while observational studies almost always leave issues of causation unsettled and subject to debate. Yet the mathematical models of statistical theory are identical for observational and experimental data.

The calcium study was in fact an experiment:

Example 2c. The design of the calcium study **[21]**. Examination of a large sample of people revealed a relationship between calcium intake and blood pressure. The relationship was strongest for black men. Researchers therefore conducted an experiment.

The subjects in part of the experiment were 21 healthy black men. A randomly chosen group of 10 of the men received a calcium supplement for 12 weeks. The control group of 11 men received a placebo pill that looked identical. The experiment was double-blind.

Can we conclude that calcium has caused a reduction in blood pressure? Such an inference, that an observed difference may be taken at face value, stands on three legs. Two of the three are grounded in data production:

(1) an argument—automatic only for random samples and randomized experiments—that a probability model applies to the data;
(2) an argument—probability-based, and comparatively straightforward—that the observed difference is "real," i.e., too big to be plausibly explained as due just to chance variation; and
(3) an argument—often thorny and fraught with pitfalls, except in the case of randomized experiments—that the observed difference is not due to some confounding influence distinct from the factor of interest.

The *t*-test of Example 2a, like all statistical tests and confidence intervals, deals only with the second argument: "If we assume that a particular chance model applies, how likely is it to get an observed difference this big?" The other two arguments depend on the design.

The clinical trial on the effect of calcium on blood pressure was a *randomized comparative experiment.* Figure 3 presents the design in outline form. The great virtue of assigning the subjects at random is that it makes arguments (1) and (3) automatic, and so reduces the problem of inferring cause to checking the fit of a model, and then, given adequate fit, carrying out a straightforward calculation. The random assignment of subjects eliminates bias in forming the treatment groups and produces groups that differ only through chance variation before we apply the treatments. The comparative design reminds us that all subjects are treated exactly alike except for the contents of the pills they take. Thus if we observe differences in the mean reduction in blood pressure greater than could be expected to arise by chance, we can be confident that the calcium brought about the effect we see.

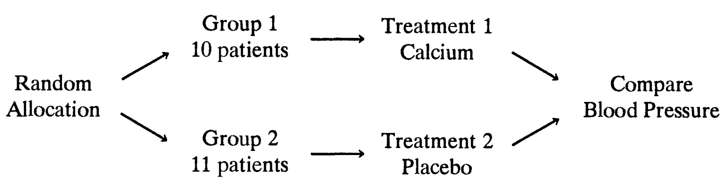


Figure 3. The simplest randomized comparative experiment.

The other major means of producing data are *sample surveys* that choose and examine a sample in order to produce information about a larger population. Interesting examples abound—opinion polls sound and unsound, government collection of economic and social data, academic data sources such as the National Opinion Research Center at the University of Chicago. Statistical designs for sampling begin by insisting that impersonal chance should choose the sample. The central idea of statistical designs for producing data, through either sampling or experimentation, is the deliberate use of chance. Explicit use of chance mechanisms eliminates some major sources of bias. It also ensures that quite simple

probability models describe our data production processes, and therefore that standard inference methods apply. However, unlike randomized experiments, observational studies do not lend themselves in so straightforward a way to an inference of causation, as the following example shows. The original study by Best and Walker appears as an example in [**12**]; our presentation here follows [**26**].

Example 3. Smoking and health. One of the early observational studies of smoking and health compared mortality rates for three groups of men. The rates, in deaths per year per 1000 men, were:

Non-smokers 20.2,   Cigarette smokers 20.5,   Cigar and pipe smokers 35.5.

To test whether the observed differences might be due to chance, we could use a model similar to the one in Example 2a. The sample sizes were so large that we can easily rule out chance variation as an explanation for the observed differences, leaving us with the apparent conclusion that cigarettes pose little risk but pipes or cigars or both are quite dangerous. Indeed, that conclusion would be valid *if* these data had come from a randomized, controlled double-blind experiment like the calcium study. However, the premise is clearly untenable. Because this is an observational study, we need to ask about other factors, linked to smoking habits, that might be responsible for the observed difference. Here, age is the main such factor: pipe and cigar smokers tend to be older than cigarette smokers, and the risk of death increases with age. In this study, the average ages for the three groups were:

Non-smokers 54.9 years,   Cigarette smokers 50.5 years,
Cigar and pipe smokers 65.9 years.

Only after adjusting the death rates for the differences in age do we get numbers more in line with what we have come to expect:

Non-smokers 20.3,   Cigarette smokers 28.3,   Cigar and pipe smokers 21.2.

Taken together, the last two examples offer what we consider two of the most important lessons for mathematicians who teach statistics: one, the conclusions from a study depend crucially on how the data were produced, and two, the standard mathematical models ignore data production.

Statistical ideas for producing data to answer specific questions are the most influential contributions of statistics to human knowledge. Badly designed data production is the most common serious flaw in statistical studies. Well designed data production allows us to apply standard methods of analysis and reach clear conclusions. Professional statisticians are paid for their expertise in designing studies; if the study is well designed (and no unanticipated disaster occurred), you don't need a professional to do the analysis. In other words, the design of data production is *really* important. If you just say "Suppose X_1 to X_n are iid observations," you aren't teaching statistics.

2.2. Data analysis: exploration and description. Data analysis is the contemporary form of "descriptive statistics," powered by more numerous and more elaborate descriptive tools, but especially by a philosophy due in large measure to John Tukey of Bell Labs and Princeton. The philosophy is captured in the now-common name, *exploratory data analysis*, or EDA. The goal of EDA is to see what the data in hand say, on the analogy of an explorer entering unknown lands. We put aside (but not forever) the issue of whether these data represent any larger universe.

Table 1 presents an elementary summary [25] of the distinctions between EDA and standard inference:

TABLE 1. EXPLORATORY DATA ANALYSIS VS. FORMAL PROBABILITY-BASED INFERENCE

Exploratory Data Analysis	**Statistical Inference**
Purpose is unrestricted exploration of the data, searching for interesting patterns.	Purpose is to answer specific questions, posted before the data were produced
Conclusions apply only to the individuals and circumstances for which we have data in hand	Conclusions apply to a larger group of individuals or a broader class of circumstances
Conclusions are informal, based on what we see in the data.	Conclusions are formal, backed by a statement of our confidence in them

In practice, exploratory analysis is a prerequisite to formal inference. Most real data contain surprises, some of which can invalidate or force modification of the inference that was planned. This is one reason why running data through a sophisticated (and therefore automated) inference procedure before exploring them carefully is the mark of a statistical novice. The dialog between data and models continues with more advanced diagnostic tools that allow data to criticize specific models. These tools combine the EDA spirit with the results of mathematical analysis of the consequences of the models.

As we have already seen, the model of Example 2a, because it does not distinguish between observation and experiment, is incomplete. It is also, like most idealized mathematical models for real phenomena, unrealistic. In the words attributed to the statistician George Box, "All models are wrong, but some are useful." The user of inference methods based on this model must carefully explore its adequacy to the setting and the data. Were there flaws in the data production (whether sample or experiment) that render inference meaningless? Are the data, which are certainly not independent observations on a perfectly normal distribution, sufficiently normal to allow use of standard procedures? This question is answered by exploratory examination of the data themselves, combined with knowledge of how "robust" the planned analysis is under deviations from the assumptions of the model.

Example 2d. Preliminary exploration of the calcium data. An analysis might start from a simple outline: plot, shape, center, spread.

Plot. A stemplot splits each data value into a stem and leaf, then sorts leaves onto shared stems. Figure 4 shows a back-to-back stemplot useful for comparing two groups:

Placebo		Calcium
1	−1	
5	−0	5
33111	−0	234
3	0	1
5	0	7
2	1	01
	1	78

Figure 4. Parallel stemplot of reduction in systolic blood pressure for two groups of men.

Shape. The distribution for the placebo group is unimodal and symmetric. The treatment group, however, contains a faint suggestion of bimodality, which raises the possibility of two kinds of subjects. Might there be some who respond to calcium, and others who do not? There is no way to tell from these data, but the possibility is worth noting.

Center and spread. A useful plot for comparing centers, spreads and symmetries is the boxplot (Figure 5). Each box locates the quartiles and median of a distribution; the "whiskers" extend from the quartile to the most extreme points within 1.5 interquartile ranges of the nearest quartile, and points at a greater distance from the median are shown separately. Here we find a difference in medians, but also a pronounced difference in spreads, one that should raise suspicions about the assumption of equal variances used to justify a pooled estimate in Example 2a.

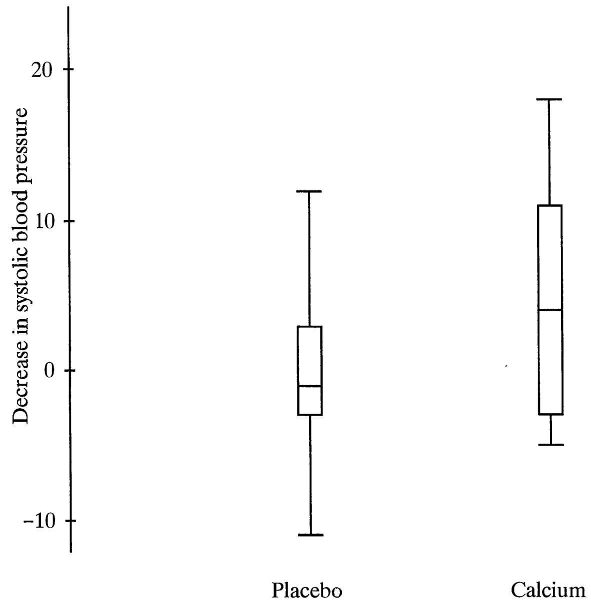


Figure 5. Parallel boxplots of reduction in systolic blood pressure for two groups of men.

Normal quantile plot. Looking ahead to a t-test to compare means, it is prudent to ask whether the data give us reason to question the normal model of Example 2a. Here we subtract the group mean from each observation to get residuals, then plot the ordered residuals against the corresponding quantiles of a normal distribution; see Figure 6. Our ordinates are the 21 ordered residuals, which divide the real line into 22 sub-intervals. The corresponding abscissas are the 21 values that divide the real line into 22 segments that are equiprobable under the normal model. If the data come from a single normal distribution, we can expect the points to fall near a line.

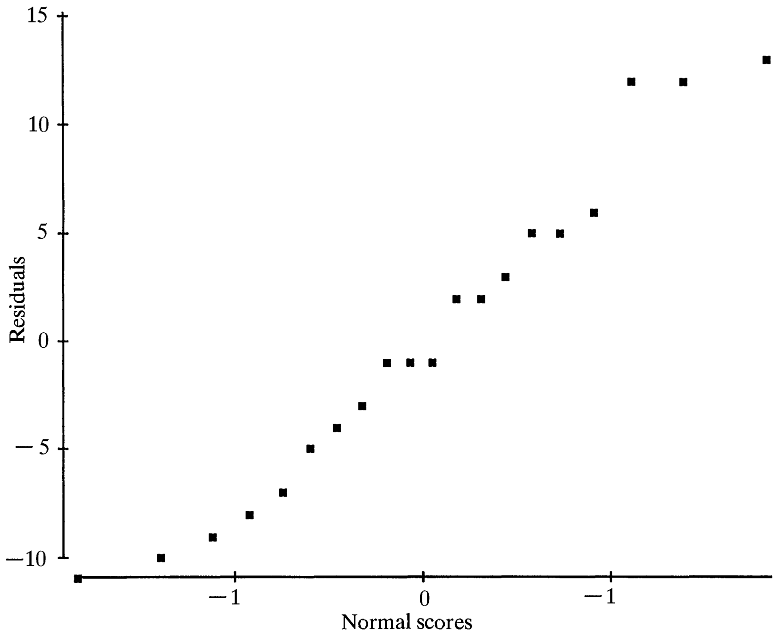


Figure 6. Normal quantile plot for the blood pressure data.

For the calcium data, the pattern is reasonably linear, although the vertical jump before the three right-most points shows observed residuals that are larger than predicted by the normal model, a pattern consistent with the unequal spreads in the boxplots.

Mathematically structured instruction, which tends to emphasize how methods follow from models, often provides only the most general warnings about the realities of practice. Statistics in practice resembles a dialog between models and data. Models for the process that produced our data do indeed play a central role in statistical inference. The mathematical exploration of properties and consequences of models is therefore important (as it is in economics and physics). But the data are also allowed to criticize and even falsify proposed models. In the calcium examples, the exploratory analysis warns us not to rely heavily on the assumption of equal variances, and to use a modified t-test that estimates separate variances for the two groups. We can modify Box's dictum into a practical version of the statement that statistics is not just mathematics: *Mathematical theorems are true*; *statistical methods are sometimes effective when used with skill.*

Wide availability of cheap computing, especially graphics, has combined with the desire to "let the data speak" to generate an abundance of new tools: at the low end we have the stemplots and boxplots of Example 2c; but there are also model-free scatterplot smoothers, resistant regression algorithms, clever ideas for display of high-dimensional data on two-dimensional screens, and many still more advanced diagnostic tools for specific situations. Standard statistical software implements much of this. The books [7] and [9], by Bell Labs scientists influenced by Tukey, present much of the basic graphical material. The software packages S and S-PLUS, which originated at Bell Labs, implement more of the new graphics and also implement several new classes of models. See [8] for detailed discussion of the latter.

Although it may be tempting for the neophyte to view data analysis as merely a collection of clever tools, the value of these tools comes from using them in a systematic way, according to strategies that organize the examining of data:

1. Proceed from simple to complex: first examine each variable individually, then look at relationships among them.
2. Use a hierarchy of tools: first plot the data, then choose appropriate numerical descriptions of specific aspects of the data, then if warranted select a compact mathematical model for the overall pattern of the data.
3. Look at both the overall pattern and at any striking deviations from that pattern.

It is part of the unifying (but non-mathematical) theory of EDA that these principles apply in each of several settings. Given data on a single quantitative variable, we might display the distribution by a stemplot, note that it reasonably symmetric, calculate the mean and standard deviation as numerical summaries, and use a normal quantile plot to see whether a normal distribution is a suitable compact model for the overall pattern. Given two quantitative variables, we draw a scatterplot, measure the direction and strength of linear association by the correlation, and, if warranted, use a fitted straight line as a model for the overall pattern. Thus the univariate "*Plot*, *shape*, *center*, *spread*," returns in the context of bivariate data as "*Plot*, *shape*, *direction*, *strength*."

Here, as elsewhere, an analysis is not just a search for patterns, but a search for *meaningful* patterns. The *best* fit is not necessarily the *most useful*, as the following example illustrates.

Example 3. Dormitories and cities. Each point in Figure 7 represents one of the 50 U.S. states with horizontal coordinate equal to the state's urban population, and

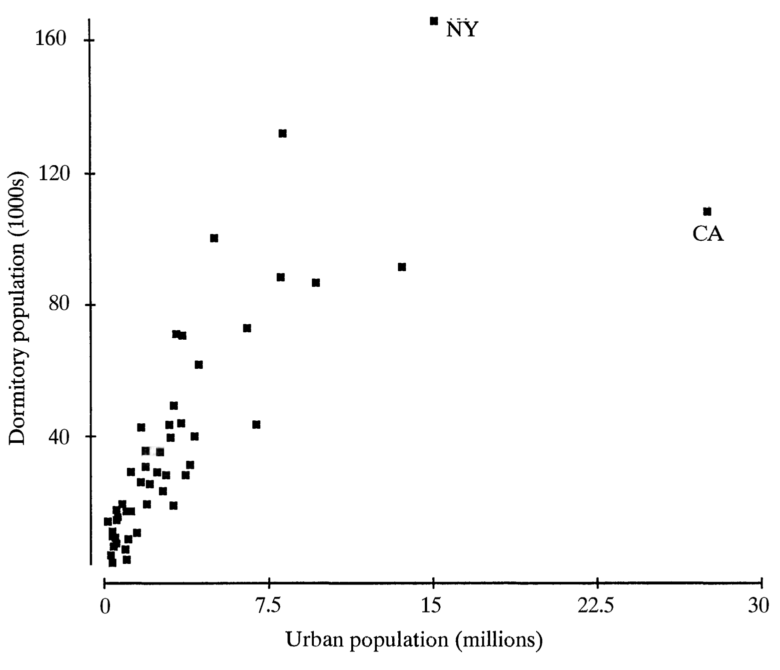


Figure 7. Scatterplot of dormitory population versus urban population for the 50 U.S. states.

vertical coordinate equal to the number of the state's college students housed in dormitories. Several features of the plot's shape stand out. For example, the plot is fan shaped, with many points bunched in the lower left: most states have relatively small urban populations (a couple of million or so) and relatively small dormitory populations as well (under 50,000); only a few states have very large urban populations or very large dormitory populations, and the variability from state to state is larger (more space between points) for the states with larger values. The pattern of association between the two variables is positive and strong: smaller urban populations go with smaller dormitory populations, larger urban populations with larger dormitory populations and, for all but a few of the states, knowing the size of a state's urban population allows us to predict its dormitory population to within a fairly narrow range.

Despite the nice fit between picture and story, the analysis so far has overlooked a most important feature. If we take at face value the pattern that states with large urban populations also have large dormitory populations, we might be tempted to conclude that cities must attract colleges. Although plenty of confirming instances come to mind, this naive interpretation is wrong: both our variables are indirect measures of the size of the states' populations, so it is hardly surprising that the two measures show a strong positive association. To uncover a more meaningful relationship, we have to "adjust for the lurking variable:" divide urban population by total population to get percent urban, divide dormitory population by total population to get percent living in dormitories, and plot the result (Figure 8).

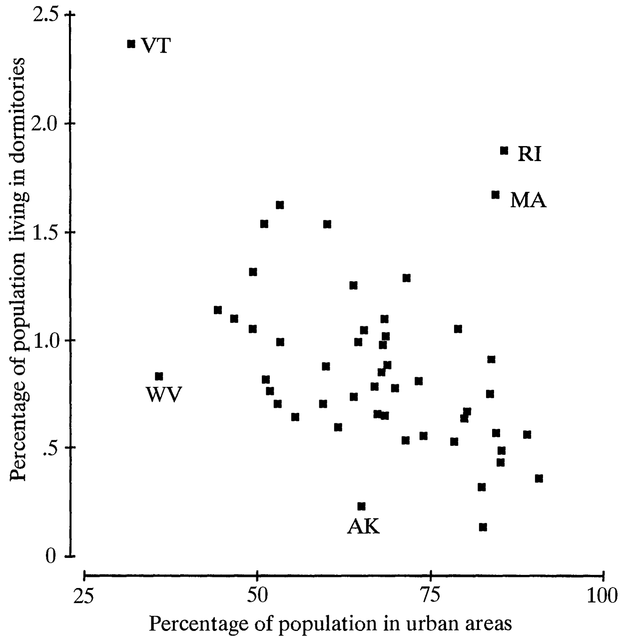


Figure 8. Scatterplot of the dorms-and-cities data after adjusting for the "lurking variable" population.

Now the relationship is weaker, but what it tells us is more interesting. The direction is reversed: rural states—those with a lower percentage of their residents living in metropolitan areas—have a higher percentage of their residents living

in college dormitories. On reflection, this makes sense. Think about Pullman, Washington, or Ames, Iowa; about Norman, Oklahoma, or Lawrence, Kansas. Rural states may have fewer colleges and universities in absolute numbers, but their students make up a higher percentage of the total population of the state, and are more likely to live in dormitories.

2.3. Formal inference: the argument against chance. Statistical inference provides methods for drawing conclusions from data about the population or process from which the data were drawn. It now becomes essential (as it was not in data analysis) to distinguish sample *statistics* from population *parameters*. The true values of the parameters are unknown to us. We have the statistics in hand, but they would take different values if we repeated out data production. Inference must take this sample variability into account.

Probability describes one kind of variability, the chance variability in random phenomena. When a chance mechanism is explicitly used to produce data, probability therefore describes the variation we expect to see in repeated samples from the same population or repeated experiments in the same setting. That is, probability answers the question, "What would happen if we did this many times?" Standard statistical inference is based on probability. It offers conclusions from data *along with* an indication of how confident we are in the conclusions. The statement of confidence is based on asking "What would happen if I used this inference method many times?" That is exactly the kind of question probability can answer, which is why we ask it. The indication of our confidence in our methods, expressed in the language of probability, is what distinguishes formal inference from informal conclusions based on, e.g., an exploratory analysis of data.

Any particular inference procedure starts with a statistic, perhaps several statistics, calculated from the sample data. The *sampling distribution* is the probability distribution that describes how this statistic would vary if we drew many samples from the same population. In elementary statistics we present two types of inference procedures, confidence intervals and significance tests. A confidence interval estimates an unknown parameter. A significance test assesses the evidence that some sought-after effect is present in the population.

A *confidence interval* consists of a recipe for estimating an unknown parameter from sample data, usually of the form "estimate $\pm$ margin of error" and a confidence level, which is the probability that the recipe actually produces an interval that contains the true value of the parameter. That is, the confidence level answers the question, "If I used this method many times, how often would it give a correct answer?"

A *significance test* starts by supposing that the sought-after effect is *not* present in the population. It asks "In that case, is the sample result surprising or not?" A probability (the p-value) says how surprising the sample result is. A result that would rarely occur if the effect we seek were absent is good evidence that the effect is in fact present. Figure 9 illustrates this reasoning in our medical example. The normal curves in that figure represent the sampling distribution of the difference $\bar{x} - \bar{y}$ between the mean blood pressure decreases in the calcium and placebo groups, for the case of no difference between the two population means. This distribution, which shows the variability due to chance alone, has mean 0. Outcomes greater than 0 come from experiments in which calcium reduces blood pressure more than the placebo. If we observe result A, we are not surprised; an outcome this far above 0 would often occur by chance. It provides no credible evidence that calcium beats the placebo. If we observe result B, on the other hand,

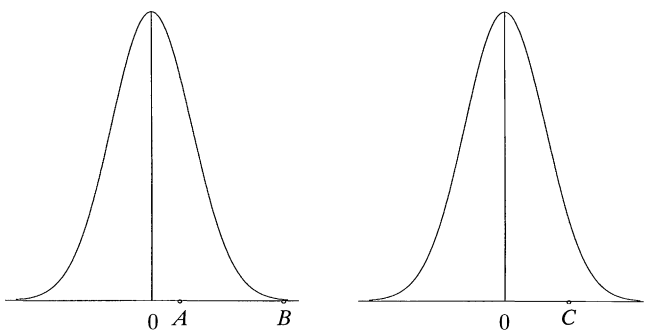


Figure 9. The idea of statistical significance: is this observation surprising?

the experiment has produced an effect so strong that it would almost never occur simply by chance. We then have strong evidence that the calcium mean does exceed the placebo mean. The p-value (the right tail probability) is 0.24 for point A and 0.0005 for point B. These probabilities quantify just how surprising an observation this large is when there is no effect in the population. What about the actual data? Point C shows the observed value $\bar{x} - \bar{y} = 5.273$. The corresponding p-value is 0.055. Calcium would beat the placebo by at least this much in 5.5% of many experiments just by chance variation. The experiment gives some evidence that calcium is effective, but not extremely strong evidence. A note for those who worry about details: These p-value calculations took the variability of the sample means to be known. In practice, we must estimate standard deviations from the data. The resulting test has a larger p-value: $p = 0.072$.

3. TEACHING. In discussing our teaching, we may focus on *content*, what we want our students to learn, or on *pedagogy*, what we do to help them learn. These two topics are of course related. In particular, changes in pedagogy are often driven in part by changing priorities for what kinds of things we want students to learn. It is nonetheless convenient to address content and pedagogy separately. This section, in keeping with the rest of this article, concerns content, and in particular contains one side of a conversation between statisticians and mathematicians who may find themselves teaching statistics.

3.1. Statistics should be taught as statistics. Statisticians are convinced that statistics, while a mathematical science, is not a subfield of mathematics. Like economics and physics, statistics makes heavy and essential use of mathematics, yet has its own territory to explore and its own core concepts to guide the exploration. Given those convictions, we would naturally prefer that beginning statistics be taught *as statistics*. The American Statistical Association and the MAA have formed a joint committee to discuss the curriculum in elementary statistics. The recommendations of that group reflect the view that statistics instruction should focus on *statistical* ideas. Here are some excerpts **[10]**; a longer discussion appears in **[11]**:

> Almost any course in statistics can be improved by more emphasis on data and concepts, at the expense of less theory and fewer recipes. To the maximum extent feasible, calculations and graphics should be automated.

> Any introductory course should take as its main goal helping students to learn the basics of statistical thinking. [These include] the need for data, the importance of data production, the omnipresence of variability, the quantification and explanation of variability.

The recommendations of the ASA/MAA committee reflect changes in the field of statistics over the past generation. Academic statistics, unlike mathematics, is linked to a larger body of non-academic professional practice. Computing technology has completely changed the practice of statistics. Academic researchers, driven in part by the demands of practice and in part by the capability of new technology, have changed their taste in research. Bootstrap methods, nonparametric data smoothing, regression diagnostics, and more general classes of models that require iterative fitting are among the recent fruits of renewed attention to analysis of data and scientific inference. Efron and Tibshirani **[14]** describe some of this work for non-specialists.

3.2. Neither Mathematics Nor Magic. An over-emphasis on probability-based inference is one mark of an overly mathematical introduction to statistics, and yet the reluctance of mathematically trained teachers to abandon a theory-driven presentation of basic statistics has a respectable basis: to avoid presenting statistics as magic. It is certainly common to teach beginning statistics as magic. The user of statistics is in many ways very like the sorcerer's apprentice. The incantation has an automatic effectiveness, rendering theses acceptable and studies publishable. We are not meant to understand how the incantation works—that is the domain of the sorcerer himself. The incantation must follow the recipe exactly, lest disaster ensue —exploration and flexibility, like understanding, are forbidden to the apprentice. Fortunately, the sorcerer has provided software that automates the exact following of approved incantations.

The danger of statistics-as-magic is real. But the proper defense is not a retreat to a mathematical presentation that is inadequate to the subject and often incomprehensible to students. *Mathematical understanding is not the only kind of understanding.* It is not even the most helpful kind in most disciplines that employ mathematics, where understanding of the target phenomena and core concepts of the discipline take precedence. We should attempt to present an intellectual framework that makes sense of the collection of tools that statisticians use and encourages their flexible application to solve problems. Students understand mathematics when they appreciate the power of abstraction, deduction, and symbolic expression, and can use mathematical tools and strategies flexibily in dealing with varied problems. Reasoning from uncertain empirical data is a similarly powerful and pervasive intellectual method. How can we best lead our students to understand, appreciate, and begin to assimilate this intellectual method?

3.3. Begin with exploratory data analysis. Although the implied chronology of Figure 2 suggests starting with data production, experience says otherwise. For one thing, exploratory data analysis makes a better beginning because it is more concrete. There is no need to distinguish population and sample, and no need to discuss the features of randomization that protect against bias. Basic methods are conceptually and algorithmically simple, and the data are in hand—actual numbers on a page, as opposed to mere ghosts of data-in-the-future, the way they are in designing an experiment. Moreover, providing motivation is not a problem. Students like exploratory analysis and find that they can do it, a substantial bonus when teaching a subject feared by many. Engaging them early on in the interpretation of results, before the harder ideas come along to claim their attention, can

help establish good habits that pay dividends when you get to inference. Finally, starting with data analysis prepares for design and for inference. Experience with data distributions introduces students to the omnipresence of variability, and to the potential for bias, the two main reasons we need careful design. If you teach design before data analysis, it is harder for students to understand why design matters. Experience with data distributions is also the best way to get ready to tackle the difficult idea of a sampling distribution.

We have tried to suggest that there is a coherent (though not mathematical) set of ideas and associated tools for exploring data. Students need to practice these ideas and tools by writing coherent descriptions of data. To help them, we provide both outlines for what to write, and examples that can serve as models. Figure 10, for example, is the outline for describing a single quantitative variable.

A. Describe the data

   number of observations
   nature of the variable
   how it was measured
   units of measurement

B. Plot the data, choose from

   dotplot
   stemplot
   histogram

C. Describe the overall pattern

   shape
      no clear shape?
      skew or symmetric?
      single or multiple peaks?

   center and spread; choose from

      five-number summary
      mean and standard deviation

   is normality an adequate model (normal quantile plot)?

D. Look for striking deviations from the overall pattern

   outliers
   gaps or clusters

E. Interpret your findings in C and D in the language of the problem setting. Suggest plausible explanations for your findings.

Figure 10. Outline for describing data on a single quantitative variable.

Following this outline requires both knowledge of the tools it mentions and judgment to choose among them and interpret the results. Judgment is formed by experience with data. Students cannot at first "read" graphs any more than they can read words or equations. Here is an example of a basic one-variable data analysis. Describing relations among several variables requires more elaborate tools and finer judgment.

> In a study of resistance to infection [**2**], researchers injected 72 guinea pigs with tubercle bacilli and measured their survival time in days after infection. Both a histogram (Figure 11) and a normal quantile plot (Figure 12) show that the distribution of survival times is strongly skewed to the right. There are no outliers—although some individuals survived far longer than the average, this appears to be a characteristic of the overall distribution rather than pointing to, for example, errors in measuring or recording these individuals.

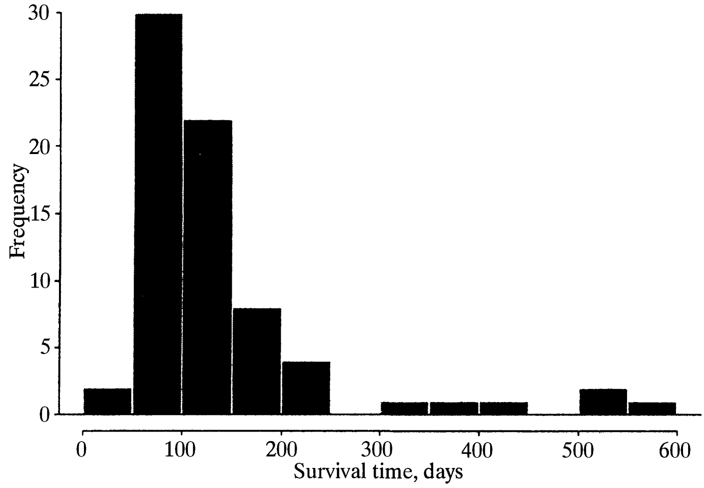


Figure 11. Histogram of guinea pig survival times.

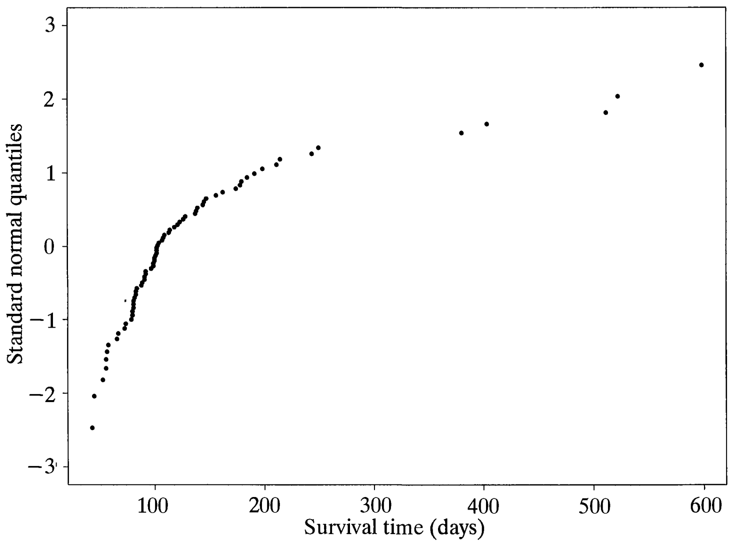


Figure 12. Normal quantile plot for guinea pig survival times.

The strong skewness suggests that the five number summary (min = 43 days, first quartile = 82.5 days, median = 102.5 days, third quartile = 151.5 days, max = 598 days) is a better numerical summary than the mean and standard deviation ($\bar{x} = 141.8$ days, $s = 109.2$ days). There is very large variation in survival times among the individuals—for example, the third quartile is almost 150% of the median and the largest 6 observations are more than double the median. Without more information, we cannot accurately predict the survival time of an infected individual. Moreover, standard t procedures should not be used for inference about survival time. Inference could employ a non-normal distribution as a model or seek a transformation to a scale that is more nearly normal.

Although many students come to a first statistics course expecting empty ritual, EDA offers them the pleasant surprise that the methods exist to serve

the search for meaning. This surprise is so welcome that it carries a danger of pushing the pendulum too far the other way. Some students may drift into a complacent conviction that any story about the data that fits the patterns with coherence and plausibility must be true. The timing is right for a dose of design and skepticism.

3.4. Teach design as the bridge between data analysis and inference. An introduction to design for data production fits naturally between exploratory analysis and inference: sound design is what makes inference possible. Waiting to introduce probability distributions until after the basics of design has a number of advantages. For one thing, this order helps make clear that the justification for probability models must come from the randomness in the data production process, and so provides some protection against unthinking adoption of probability models. For another, learning about data production introduces students to essential concepts like population and sample, parameter and statistic, before they encounter the sampling distribution, which is conceptually difficult all by itself.

The single most important point for students to understand is why randomized comparative experiments are the gold standard for evidence of causation. A rich source of true-life cautionary tales is the book [**6**], edited by the physicians Bunker and Barnes and the statistician Mosteller, which contains striking examples of medical treatments that became standard in the days before medicine adopted randomized comparative experiments, and were found to be worthless when subjected to proper testing.

There is of course more to the statistical side of designing experiments and sample surveys than "randomize." The designs used in practice are often quite complex, and must balance efficiency with the need for information of varying precision about many factors and their interactions. Simple designs—randomized experiments comparing two or several treatments, simple random samples from one or several populations—illustrate the most important ideas and support the inference taught in a first statistics course. You must talk about these designs, but need not go farther. Some other important material, for example, procedures for developing and testing survey questions and for training and supervising interviewers, is not usually presented in statistics courses. Statistics students should be aware that these practical skills do matter, and that data production can go awry even when we start with a sound statistical design. How much time to spend here is a matter of your judgment of the needs of your audience.

3.5. Inference: two barriers to understanding. Section 2.3 has described briefly how inference works. Because the details are in practice automated, we would like students to put most of their effort into grasping the ideas. They are not easy to grasp. The first barrier is the notion of a sampling distribution. Choose a simple setting, such as using the proportion $\hat{p}$ of a sample of workers who are unemployed to estimate the proportion p of unemployed workers in an entire population. Physical examples (sampling beads from a box), computer simulations, and encouraging thought experiments all help convey the idea of many samples with many values of $\hat{p}$. Keep asking, "What would happen if I did this many times?" That question is the key to the logic of standard statistical inference.

Once the idea of a sampling distribution begins to settle, the tools of data analysis help us take the next steps. Faced with any distribution, we ask about shape, center, and spread. The shape of the sampling distribution of $\hat{p}$ is approximately normal. The mean is equal to the unknown population proportion p. This says that $\hat{p}$ as an estimator of p has no bias, or systematic error. The precision of

the estimator is described by the spread of the sampling distribution, which (thanks to normality) we measure by its standard deviation. We are now only details away from confidence intervals.

The second major barrier is the reasoning of significance tests. Although the basic idea ("Is this outcome surprising?") is not recondite, the details are daunting. There's no escape from null and alternative hypotheses and one- versus two-sided tests. The logic of testing, which starts out "Suppose for the sake of argument that the effect we seek is not present ... " isn't straightforward. We'd like most of our students to understand the idea of a sampling distribution; we know that quite a few won't understand the reasoning of significance tests. Our fallback position is to insist that they be able to verbalize the meaning of p-values produced by software or reported in a journal. This is part of insisting that students write succinct summaries of statistical findings. "The study compared two methods of teaching reading to third-grade students. A two-sample t test comparing the mean scores of the two treatment groups on a standard reading test had p-value $p = 0.019$. That is, the study observed an effect so large that it would occur just by chance only about 2% of the time. This is quite strong evidence that the new method does result in a higher mean score than the standard method."

Two concluding remarks about inference. First, a conceptual grasp of the ideas is almost pictorial, based on picturing the sampling distribution and following the tactics learned in data analysis. No amount of formal mathematics can replace this pictorial vision, and no amount of mathematical derivation will help most of our students see the vision. The mathematics is essential to our knowing the facts, but this does not imply that we should impose the mathematics on our students.

Second, we want our students to know a good deal more than the big picture and several recipes that implement it in specific settings. Here are some further points, both practical and conceptual, roughly in order of importance. How far down the list you should go depends on your audience.

- Study of specific inference procedures reveals behaviors that are common and that all students should understand. To get higher confidence from the same data, you must pay with a larger margin of error. Even effects so small as to be practically unimportant are highly significant in the statistical sense if we base a significance test on a very large sample.
- Lots of things can go wrong that make inference of dubious value. Comparing subjects who *choose* to take calcium against others who don't tells little about the effects of calcium, because those who choose to take calcium may be very health-conscious in general. One extreme outlier could pull the conclusion of our medical experiment in either direction, again invalidating the inference. Examine the data production. Plot the data. Then, perhaps, go on to inference.
- Inference procedures themselves don't tell us that something went wrong. The margin of error in a confidence interval, for example, includes *only* the chance variation in random sampling. As the *New York Times* says in the box that accompanies its opinion poll results, "In addition to sampling error, the practical difficulties of conducting any survey of public opinion may introduce other sources of error into the poll."
- Common inference procedures really are based on mathematical models like the one that appears in our medical example: $X_1, X_2, \dots, X_n$ iid $N(\mu_1, \sigma_1)$, $Y_1, Y_2, \dots, Y_m$ iid $N(\mu_2, \sigma_2)$. This model isn't exactly true; is it useful? In fact, the two-sample t procedures that follow from this model when we want to

compare μ_1 and μ_2 are quite robust against non-normality, so the model does lead to practically useful procedures. But the variance ratio F statistic for comparing σ_1 and σ_2 is extremely sensitive to non-normality, so much so that it is of little practical value. Even beginners need to be aware of such issues.

- We often want to do inference when our data do not come from a random sample or randomized comparative experiment. Think, for example, of measurements on successive parts flowing from an assembly line. Inference is justified by a probability model for the process that produced our data, and the correctness of the model can to some extent be assessed from the data themselves. Randomized data production is the paradigm and the most secure setting for inference, but it is not the only allowable setting.
- Inductive inference from data is conceptually complex. It's not surprising that there are alternative ways of thinking about it. Standard statistical theory tends to think of inference as if its purpose were to make decisions. A test must decide between the null and alternative hypotheses, for example. This leads at once to Type I and Type II errors and so on. The decision-making approach fits uneasily with the "Is this outcome surprising?" logic expressed by p-values. We think that assessing the strength of evidence is a much more common goal than making a decision, but not everyone agrees. The Bayesian school of thought goes farther, by introducing an explicit description of the available prior information into any statistical setting and combining prior information with data to reach a decision. Almost all statisticians think this is sometimes a good idea. Bayesians think *all* statistical problems can be made to fit this paradigm. This is a (strongly held) minority position. Deep water ahead.

3.6. What About Probability? Probability is an essential part of any mathematical education. It is an elegant and powerful field of mathematics that enriches the subject as a whole by its interactions with other fields of mathematics. Probability is also essential to serious study of applied mathematics and mathematical modeling. The domain of determinism in natural and social phenomena is limited, so that the mathematical description of random behavior must play a large role in describing the world. Whether our mathematical tastes run to purity or modeling, probability helps to satisfy them. Here, however, we are discussing introductory statistics rather than mathematics.

From the point of view of deductive logic that has shaped so much of statistical teaching in the past, probability is more basic than statistics: probability provides the chance models that describe the variability in observed data. From the point of view of the development of understanding, however, we believe that statistics is more basic than probability: whereas variability in data can be perceived directly, chance models can be perceived only after we have constructed them in our own minds. In the ideal Platonic world of mathematics, we can start with a probabilistic chicken and use deductive logic to lay a statistical egg, but in the messier world of empirical science, we must start with the egg as observed data and construct a prior probabilistic chicken as an inference. In an introductory *statistics* course, the chicken's only value is to explain where eggs come from. It seems a bit unfair, in that context, at least, to ask beginning students to learn about egg-generators before they've become familiar with eggs—less extreme, but in the same spirit as starting the study of chemistry with quantum mechanics.

What then, should be the place of probability in beginning instruction in statistics? Our position is not standard, though it is gaining adherents: first courses in statistics should contain essentially no formal probability theory.

Why? First, because *informal probability is sufficient for a conceptual grasp of inference*. Although the theoretical structure of standard statistical inference is based on probability, the role of probability is limited to answering the question "What would happen if we used this method very many times?" The answer is given by the sampling distribution of a statistic, which records the pattern of variation of the outcomes of, for example, many random samples from the same population. If we agree that actually deriving these distributions is better left to more advanced study, they can be understood as distributions using the tools of data analysis, without the apparatus of formal probability. Rules for $P(A \cup B)$ add very little to a statistics course.

The second reason to avoid formal probability is that *probability is conceptually the hardest subject in elementary mathematics*. The history of probabilistic ideas (see **[16]** and **[27]**) is fascinating but a bit frightening. Better minds than ours long found the subject confusing in the extreme. Psychologists, beginning with Tversky and his collaborators, have demonstrated that confusion persists, even among those who can recite the axioms of formal probability and who can do textbook exercises. Our intuition of random behavior is gravely and systematically defective; see, e.g., **[28]** and the collection **[19]**. What is worse, mathematics educators have found no effective way to correct our defective intuition. Garfield and Ahlgren **[15]** conclude a review of research by stating that "teaching a conceptual grasp of probability still appears to be a very difficult task, fraught with ambiguity and illusion." They suggest study of "how useful ideas of statistical inference can be taught independently of technically correct probability." We believe that concentrating on the idea of a sampling distribution allows this, at least at the depth appropriate for beginners.

The concepts of statistical inference, starting with sampling distributions, are of course also quite tough. We ought to concentrate our attention, and ours students' limited patience with hard ideas, on the essential ideas of statistics. We faculty imagine that formal probability illumines those ideas. That's simply not true for almost all of our students.

3.7. What About Mathematics Majors? Mathematics majors traditionally meet statistics as the second course in a year-long sequence devoted to probability and statistical theory. We hope it is clear that we don't regard a tour of sufficient statistics, unbiasedness, maximum likelihood estimators, and the Neyman-Pearson theorem as a promising way to help students understand the core ideas of statistics. On the other hand, mathematics majors should certainly see some of the mathematical structure of statistical inference. What ought we do?

Our preference is to precede the study of theory by a thorough data-oriented introduction to statistical ideas and methods and their applications. That is, mathematics students are not necessarily an exception to the principle that a first introduction to statistics should not be based on formal probability. If the students have strong quantitative backgrounds, a data-oriented course can move quickly enough to present genuinely useful statistics and serious applications. The need for theory can be made clear as we face issues of practice, and the theory makes much more sense when its setting in practice is clear. In many institutions, however, constraints or faculty hesitation make this path difficult. In others, there is little coordination between the "applied" and theoretical courses, so that the latter does not in fact build on the former.

We ought therefore to reconsider what a one-semester introduction to statistics for mathematics majors and other quantitatively strong students should look like.

This course would ordinarily and most easily follow a course in probability. Here we encounter another barrier: we can't in good conscience retool both semesters of the standard probability-statistics sequence to optimize the introduction to statistics. Probability is important in its own right, not just as preparation for statistical theory. The more emphasis a department places on applications and modeling in its major curriculum, the more the probability course must play an essential role in this emphasis. An introduction to probability that emphasizes modeling and includes simulation and numerical calculation certainly sets the stage for statistics, but we are hesitant to move any strictly statistical ideas into the probability semester. The reform of probability and the reform of statistics are distinct issues.

Our goal should be an integrated statistics course that moves through data analysis, data production, and inference in turn, emphasizing the organizing principles of each. We should certainly take advantage of and strengthen the student's mathematical capacities. Although data analysis and data production have no unifying theory, mathematical analysis can illumine even data analysis. Here are a few examples.

- A. Consider the optimality properties of measures of center for n observations. The mean minimizes the mean squared error; the median minimizes the mean absolute error (and need not be unique); the midrange minimizes the maximum absolute (or squared) error; try minimizing the *median* absolute error for $n = 3$ and examine the unpleasant behavior of the resulting measure.
- B. Students met the Chebychev inequality while studying probability. Now they may meet the interesting inequality $|\mu - m| \leq \sigma$ linking the mean, median, and standard deviation of any distribution **[29]**. Describe one-sample data by the empirical distribution (probability $1/n$ on each observed point) to draw conclusions about how far apart the sample mean and median may be.
- C. The least-squares regression line is the analog of the mean $\bar{x}$ for predicting y from x. Derive it. Then explore, perhaps using software, analogs of the other measures mentioned in A.

Data production lends itself to probability calculations that illustrate how likely it is that random assignments will be unbalanced in specific ways; the advantages of large samples soon become clear.

Very nice. We can give our students a balanced introduction to statistics that makes use of their knowledge of mathematics. The inevitable consequence is that we spend less time on inference. We must decide what to preserve and what to cut. There is as yet no consensus, because, despite much grumbling, the reform of the math major sequence has not yet begun. Imagining such a reform is a good place to end a discussion of statistics, mathematics, and teaching. This is your take-home exam: design a better one-semester statistics course for mathematics majors.

REFERENCES

1. Aldous, David (1994), Triangulating the circle, at random, *Amer. Math. Monthly* **101**, 223–233. The remark appears in the biographical note accompanying the paper.
2. Bjerkedal, T. (1960), Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli, *American Journal of Hygiene* **72**, 130–148.
3. Boyer, Paul and Stephen Nissenbaum (1972). *Salem Village Witchcraft*. Belmont, CA: Wadsworth Publishing Co.
4. Boyer, Paul and Stephen Nissenbaum (1974). *Salem Possessed*. Cambridge, MA: Harvard University Press.

5. Bullock, James O. (1994), Literacy in the language of mathematics, *Amer. Math. Monthly* **101**, 735–743.
6. Bunker, John P., Benjamin A. Barnes, and Frederick Mosteller (eds.) (1977), *Costs, Risks and Benefits of Surgery*. New York: Oxford University Press.
7. Chambers, John M., William S. Cleveland, Beat Kleiner, and Paul A. Tukey (1983), *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.
8. Chambers, John M. and Trevor J. Hastie (1992), *Statistical Model in S*. Pacific Grove, CA: Wadsworth.
9. Cleveland, William S. and Mary E. McGill (eds.) (1988), *Dynamic Graphics for Statistics*. Belmont, CA: Wadsworth.
10. Cobb, George W. (1991), Teaching statistics: more data, less lecturing, *Amstat News*, December 1991, pp. 1, 4.
11. Cobb, George W. (1992), Teaching statistics, in L. A. Steen (ed.) *Heeding the Call for Change: Suggestions for Curricular Action*, MAA Notes 22. Washington, DC: Mathematical Association of America.
12. Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies, *Biometrics* **24**, 205–213.
13. Crystal, David (ed.) (1994), *The Cambridge Factfinder*. Cambridge: Cambridge University Press, pp. 174–175.
14. Efron, Bradley and Rob Tibshirani (1991), Statistical data analysis in the computer age, *Science* **253**, 390–395.
15. Garfield, Joan and Andrew Ahlgren (1988), Difficulties in learning basic concepts in probability and statistics: implications for research, *Journal for Research in Mathematics Education* **19**, 44–63.
16. Gigerenzer, G., Z. Swijtink, T. Porter, L. Daston, J. Beatty, and L. Krüger (1989) *The Empire of Chance*. Cambridge: Cambridge University Press.
17. Hoaglin, D. C. (1992), Diagnostics, in D. C. Hoaglin and D. S. Moore (eds.), *Perspectives on Contemporary Statistics*, MAA Notes 21. Washington, DC: Mathematical Association of America, pp. 123–144.
18. Hoaglin, David C. and David S. Moore (eds.) (1992), *Perspectives on Contemporary Statistics*, MAA Notes 21. Washington, DC: Mathematical Association of America.
19. Kapadia, R. and M. Borovcnik (eds.) (1991), *Chance Encounters: Probability in Education*. Dordrecht: Kluwer.
20. Longfellow, Henry Wadsworth (1847), *Evangeline*, Introduction, 1.1.
21. Lyle, Roseann M. et al. (1987), Blood pressure and metabolic effects of calcium supplementation in normotensive white and black men, *Journal of the American Medical Association* **257**, 1772–1776. Dr. Lyle provided the data in the example.
22. Moore, David S. (1988), Should mathematicians teach statistics (with discussion), *College Math. Journal* **19**, 3–7.
23. Moore, David S. (1992), What is statistics? in David C. Hoaglin and David S. Moore (eds.), *Perspectives on Contemporary Statistics*, MAA Notes 21. Washington, DC: Mathematical Association of America, pp. 1–18.
24. Moore, David S. (1992), Teaching statistics as a respectable subject, in Florence Gordon and Sheldon Gordon (eds.), *Statistics for the Twenty-First Century*, MAA Notes 26. Washington, DC: Mathematical Association of America.
25. Moore, David S. (1995), *The Basic Practice of Statistics*. New York: W. H. Freeman.
26. Rosenbaum, Paul R. (1995), *Observational Studies*. New York: Springer-Verlag, p. 60.
27. Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Mass: Belknap.
28. Tversky, Amos and Daniel Kahneman (1983), Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment, *Psychological Review* **90**, 293–315.
29. Watson, G. S. (1994), letter to the editor, *The American Statistician* **48**, p. 269. This is the last in a sequence of comments on this inequality, and contains references to the earlier contributions.
30. Weisberg, Sanford (1985). *Applied Linear Regression*, 2nd edition. New York: John Wiley and Sons, p. 230.

Department of Mathematics, Statistics and Computer Science
Mount Holyoke College
South Hadley, MA 01075
gcobb@mtholyoke.edu

Department of Statistics
Purdue University
West Lafayette, IN 47907
dsm@stat.purdue.edu

When Is A Linear Operator Diagonalizable?

Marco Abate

INTRODUCTION. As it often happens, everything began with a mistake. I was teaching for the third year in a row a linear algebra course to engineering freshmen. One of the highlights of the course was eigenvector theory, and in particular the diagonalization of linear operators on finite-dimensional vector spaces (i.e., of square real or complex matrices). Toward the end of the course I assigned a standard homework: prove that the matrix

$$A = \begin{vmatrix} -1 & -1 & 2 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{vmatrix},$$

is diagonalizable. Easy enough, I thought. The characteristic polynomial is

$$p_A(\lambda) = \det(A - \lambda I_3) = -\lambda^3 + \lambda,$$

whose roots are evidently $0, 1, -1$. We have three distinct eigenvalues in a three-dimensional space, and a standard theorem ensures that A is diagonalizable.

To my surprise, the students came complaining that they were unable to solve the exercise. Perplexed (some of the complaining students were very bright), I looked over the exercise again—and I understood. What happened was that, in the homework, I *actually* gave them the matrix

$$B = \begin{vmatrix} -1 & -1 & 2 \\ -1 & 0 & 1 \\ 0 & -1 & -1 \end{vmatrix},$$

whose characteristic polynomial is

$$p_B(\lambda) = -\lambda^3 - 2\lambda^2 - \lambda + 2,$$

which has *no* rational roots. The students were unable to compute the eigenvalues of B, and they got stuck.

This accident started me wondering whether it might be possible to decide when a linear operator T on a finite-dimensional real or complex vector space is diagonalizable *without* computing the eigenvalues. If one is looking for an orthonormal basis of eigenvectors, the answer is well known to be yes: the spectral theorem says that such a basis exists in the complex case if and only if T is normal (i.e., it commutes with its adjoint), and if and only if T is symmetric in the real case. The aim of this note is to give an explicit procedure to decide whether a given linear operator on a finite-dimensional real or complex vector space is diagonalizable. By "explicit" I mean that it can always be worked out with pen and paper; it can be long, it can be tedious, but it can be done. Its ingredients (the minimal polynomial and Sturm's theorem) are not new; but putting them together yields a result that can be useful as an aside in linear algebra classes.

1. THE MINIMAL POLYNOMIAL. The first main ingredient in our procedure is the minimal polynomial. Let $T\colon V \to V$ be a linear operator on a finite-dimensional vector space over the field $\mathbb{K}$. We denote by T^k the composition of T with itself k times, and for any polynomial $p(t) = a_k t^k + \cdots + a_0 \in \mathbb{K}[t]$ we put

$$p(T) = a_k T^k + \cdots + a_1 T + \mathrm{id}_V,$$

and say that p is *monic* if $a_k = 1$. A *minimal polynomial* $\mu_T \in \mathbb{K}[t]$ of the linear operator T is a monic polynomial of minimal degree such that $\mu(T) = 0$.

The theory of the minimal polynomial is standard. For completeness, I briefly recall the results we shall need. First of all:

Proposition 1.1. *Let $T\colon V \to V$ be a linear operator on a finite-dimensional vector space V over the field $\mathbb{K}$. Then*:

(i) *the minimal polynomial μ_T of T exists, has degree at most $n = \dim V$, and is unique*;
(ii) *if $p \in \mathbb{K}[t]$ is such that $p(T) = 0$, then there is some $q \in \mathbb{K}[t]$ such that $p = q\mu_T$.*

For our procedure it is important to show that the minimal polynomial can be explicitly computed. Take $v \in V$, and let d be the minimal non-negative integer such that the vectors $\{v, T(v), \dots, T^d(v)\}$ are linearly dependent. Clearly $d \le n$ always; $d = 0$ if and only if $v = 0$, and $d = 1$ if and only if v is an eigenvector of T. Choose $a_0, \dots, a_{d-1} \in \mathbb{K}$ such that

$$T^d(v) + a_{d-1}T^{d-1}(v) + \cdots + a_1 T(v) + a_0 v = 0$$

(note that we can assume the coefficient of $T^d(v)$ to be 1 because of the minimality of d), and then set

$$\mu_{T,v}(t) = t^d + a_{d-1}t^{d-1} + \cdots + a_1 t + a_0 \in \mathbb{K}[t].$$

By definition, $v \in \operatorname{Ker} \mu_{T,v}(T)$; more precisely, $\mu_{T,v}$ is the monic polynomial $p \in \mathbb{K}[t]$ of least degree such that $v \in \operatorname{Ker} p(T)$.

Now, if $p \in \mathbb{K}[t]$ is any common multiple of μ_{T,v_1} and μ_{T,v_2} for any two vectors v_1 and v_2, then both v_1 and v_2 belong to $\operatorname{Ker} p(T)$. More generally, if $\mathscr{B} = \{v_1, \dots, v_n\}$ is a basis of V, and p is any common multiple of $\mu_{T,v_1}, \dots, \mu_{T,v_n}$, then $\mathscr{B} \subset \operatorname{Ker} p(T)$, and thus $p(T) = 0$. Hence the following result comes as no surprise:

Proposition 1.2. *Let $T\colon V \to V$ be a linear operator on a finite-dimensional vector space V over the field $\mathbb{K}$. Let $\mathscr{B} = \{v_1, \dots, v_n\}$ be a basis of V. Then μ_T is the least common multiple of $\mu_{T,v_1}, \dots, \mu_{T,v_n}$.*

Proof: Let $p \in \mathbb{K}[t]$ be the least common multiple of $\mu_{T,v_1}, \dots, \mu_{T,v_n}$. We have already remarked that $p(T) = 0$, and so μ_T divides p. Conversely, for $j = 1, \dots, n$ write $\mu_T = q_j \mu_{T,v_j} + r_j$, with $\deg r_j < \deg \mu_{T,v_j}$. Then

$$0 = \mu_T(T)v_j = q_j(T)\big(\mu_{T,v_j}(T)v_j\big) + r_j(T)v_j = r_j(T)v_j,$$

and the minimality of the degree of μ_{T,v_j} forces $r_j \equiv 0$. Since every μ_{T,v_j} divides μ_T, their least common multiple p also divides μ_T, and hence $p = \mu_T$. ■

Thus one method to compute the minimal polynomial is to compute the polynomials $\mu_{T,v_1}, \dots, \mu_{T,v_n}$ and then find their least common multiple. To avoid unnecessary calculations, it could be useful to remember that $\deg \mu_T \le n$.

Example 1. Let us compute the minimal polynomials of the matrices A and B of the introduction. Let $\mathscr{B} = \{e_1, e_2, e_3\}$ be the canonical basis of $\mathbb{R}^3$. We have

$$Ae_1 = Be_1 = \begin{vmatrix} -1 \\ -1 \\ 0 \end{vmatrix}, \qquad A^2e_1 = B^2e_1 = \begin{vmatrix} 2 \\ 1 \\ 1 \end{vmatrix},$$

$$A^3e_1 = \begin{vmatrix} -1 \\ -1 \\ 0 \end{vmatrix} = Ae_1, \qquad B^3e_1 = \begin{vmatrix} -1 \\ -1 \\ -2 \end{vmatrix} = -2B^2e_1 - Be_1 + 2e_1;$$

therefore

$$\mu_{A,e_1}(t) = t^3 - t, \qquad \mu_{B,e_1}(t) = t^3 + 2t^2 + t - 2.$$

Since $\deg \mu_{A,e_1} = 3$ and the minimal polynomial of A should be a monic multiple of μ_{A,e_1} of degree at most three, we can conclude that $\mu_A = \mu_{A,e_1}$ without computing μ_{A,e_2} and μ_{A,e_3} (and it is easy to check that $\mu_{A,e_2}(t) = t^2 - t$ and $\mu_{A,e_3}(t) = t^3 - t$). For the same reason we have $\mu_B = \mu_{B,e_1}$.

Let $\lambda_1, \ldots, \lambda_k \in \mathbb{K}$ be the distinct eigenvalues of T. If T is diagonalizable, then Proposition 1.2 immediately yields $\mu_T(t) = (t - \lambda_1) \cdots (t - \lambda_k)$. This is the standard characterization of diagonalizable linear operators:

Theorem 1.3. *Let $T\colon V \to V$ be a linear operator on a finite-dimensional vector space V over the field $\mathbb{K}$. Then T is diagonalizable if and only if μ_T is of the form*

$$\mu_T(t) = (t - \lambda_1) \cdots (t - \lambda_k), \tag{1.1}$$

where $\lambda_1, \ldots, \lambda_k$ are distinct elements of $\mathbb{K}$.

Therefore to decide whether a given linear operator on a finite-dimensional vector space is diagonalizable it suffices to check whether its minimal polynomial is of the form (1.1).

2. THE PROCEDURE. Our aim now is to find an effective procedure to decide whether a given polynomial $p \in \mathbb{K}[t]$ can be written in the form (1.1). To do so, we need to know when all the roots of p have multiplicity one, and when they all belong to the field $\mathbb{K}$. The first question has a standard answer:

Proposition 2.1. *Let $p \in \mathbb{K}[t]$ be a non-constant polynomial, and let $p' \in \mathbb{K}[t]$ denote its derivative. Then the following assertions are equivalent:*

(i) *p admits a root in $\mathbb{K}$ of multiplicity greater than* 1;
(ii) *p and p' have a common root in $\mathbb{K}$;*
(iii) *the greatest common divisor* g.c.d.(p, p') *of p and p' has a root in $\mathbb{K}$.*

Recalling Theorem 1.3 we get the following

Corollary 2.2. *Let $T\colon V \to V$ be a linear operator on a finite-dimensional vector space V over the field $\mathbb{K}$. Then:*

(i) *If $\mathbb{K}$ is algebraically closed (e.g., $\mathbb{K} = \mathbb{C}$), then T is diagonalizable if and only if* g.c.d.$(\mu_T, \mu'_T) = 1$;
(ii) *If $\mathbb{K}$ is not algebraically closed (e.g., $\mathbb{K} = \mathbb{R}$), then T is diagonalizable if and only if all the roots of μ_T are in $\mathbb{K}$ and* g.c.d.$(\mu_T, \mu'_T) = 1$.

To decide whether a complex linear operator T is diagonalizable it then suffices to compute the greatest common divisor of μ_T and μ'_T. On the other hand, if $\mathbb{K} = \mathbb{R}$ this is not enough; to complete the picture we need Sturm's theorem—and to state it we need a few more definitions.

Let $\mathbf{c} = (c_0, \dots, c_s) \in \mathbb{R}^{s+1}$ be a finite sequence of real numbers. If $c_0 \cdots c_s \neq 0$, the *number of variations in sign* of $\mathbf{c}$ is the number of indices $1 \leq j \leq s$ such that $c_{j-1}c_j < 0$ (that is, such that c_{j-1} and c_j have opposite sign). If some element of $\mathbf{c}$ is zero, then the *number of variations in sign* of $\mathbf{c}$ is the number of variations in sign of the sequence of non-zero elements of $\mathbf{c}$. We denote the number of variations in sign of $\mathbf{c}$ by $V_{\mathbf{c}}$.

Now let $p \in \mathbb{R}[t]$ be a non-constant polynomial. The *standard sequence* associated with p is the sequence $p_0, \dots, p_s \in \mathbb{R}[t]$ defined by

$$\begin{aligned}
p_0 &= p, & p_1 &= p',\\
p_0 &= q_1 p_1 - p_2, & &\text{with } \deg p_2 < \deg p_1,\\
&\vdots & &\vdots\\
p_{j-1} &= q_j p_j - p_{j+1}, & &\text{with } \deg p_{j+1} < \deg p_j,\\
&\vdots & &\vdots\\
p_{s-1} &= q_s p_s, & &\text{that is, } p_{s+1} \equiv 0.
\end{aligned}$$

In other words, the standard sequence is obtained by changing the sign in the remainder term of the Euclidean algorithm for the computation of g.c.d.(p, p'). In particular, g.c.d.$(p, p') = 1$ if and only if p_s is constant.

Sturm's theorem then says:

Theorem 2.3. *Let $p \in \mathbb{R}[t]$ be a polynomial such that* g.c.d.$(p, p') = 1$, *and take $a < b$ such that $p(a)p(b) \neq 0$. Let $p_0, \dots, p_s \in \mathbb{R}[t]$ be the standard sequence associated with p. Then the number of roots of p in $[a, b]$ is equal to $V_{\mathbf{a}} - V_{\mathbf{b}}$, where $\mathbf{a} = (p_0(a), \dots, p_s(a))$ and $\mathbf{b} = (p_0(b), \dots, p_s(b))$.*

For a proof see [1, pp. 295–299].

Now, for any polynomial $p(t) = a_d t^d + \cdots + a_0 \in \mathbb{R}[t]$ there exists $M > 0$ such that $p(t)$ has the same sign as a_d, the *leading coefficient* of p, if $t \geq M$ and the same sign as $(-1)^d a_d$ if $t \leq -M$. In particular, all the roots of p are contained in $[-M, M]$, and Sturm's theorem implies the following:

Corollary 2.4. *Let $p \in \mathbb{R}[t]$ be a non-constant polynomial such that* g.c.d.$(p, p') = 1$. *Let $p_0, \dots, p_s \in \mathbb{R}[t]$ be the standard sequence associated with p, and let d_j be the degree and $c_j \in \mathbb{R}$ the leading coefficient of p_j for $j = 0, \dots, s$. Then the number of real roots of p is given by $V_- - V_+$, where V_- is the number of variations in sign of the sequence $((-1)^{d_0}c_0, \dots, (-1)^{d_s}c_s)$, and V_+ is the number of variations in sign of the sequence $(c_0, \dots, c_s)$.*

Proof: It suffices to choose $M > 0$ large enough so that $p_j(t)$ has the same sign as c_j when $t \geq M$ and the same sign as $(-1)^{d_j}c_j$ when $t \leq -M$, for each $j = 0, \dots, s$, and then apply Sturm's theorem with $a = -M$ and $b = M$. ∎

We finally have all the ingredients necessary to state the desired procedure. Let $T: V \to V$ be a linear operator on a finite-dimensional vector space V over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. Then:

(1) Compute the minimal polynomial μ_T.
(2) Compute the standard sequence $p_0, \ldots, p_s$ associated with μ_T. If p_s is not constant, then T is not diagonalizable. If p_s is constant and $\mathbb{K} = \mathbb{C}$, then T is diagonalizable. If p_s is constant and $\mathbb{K} = \mathbb{R}$, go to Step (3).
(3) Compute V_- and V_+ for μ_T. Then T is diagonalizable if and only if $V_- - V_+ = \deg\ \mu_T$.

Thus we are always able to decide whether a given linear operator on a finite-dimensional real or complex vector space is diagonalizable or not. One feature that I find particularly interesting in this procedure is that the solution of a typical linear algebra problem is reduced to an apparently totally unrelated manipulation of polynomials, showing in a simple case how different parts of mathematics can be connected in unexpected ways.

We end this note with some examples of application of our procedure.

Example 2. First of all, we solve the original homework. We have already computed the minimal polynomial $\mu_B(t) = t^3 + 2t^2 + t - 2$. The standard sequence associated with μ_B is

$$p_0(t) = t^3 + 2t^2 + t - 2, \quad p_1(t) = 3t^2 + 4t + 1,$$
$$p_2(t) = \tfrac{2}{9}t + \tfrac{20}{9}, \quad p_3(t) = -261.$$

Since p_3 is constant, B is diagonalizable over $\mathbb{C}$. To compute $V_- - V_+$ we count the number of variations in sign of the sequences $(-1, 3, -\frac{2}{9}, -261)$ and $(1, 3, \frac{2}{9}, -261)$. We obtain

$$V_- - V_+ = 2 - 1 = 1 < 3 = \deg\ \mu_B,$$

and so B is not diagonalizable over $\mathbb{R}$. On the other hand, the standard sequence associated with μ_A is

$$p_0(t) = t^3 - t, \quad p_1(t) = 3t^2 - 1, \quad p_2(t) = \tfrac{2}{3}t, \quad p_3(t) = 1.$$

The number of variations in sign of $(-1, 3, -\frac{2}{3}, 1)$ is $V_- = 3$, and of $(1, 3\,\frac{2}{3}, 1)$ is $V_+ = 0$; therefore $V_- - V_+ = 3 - 0 = 3$, and thus A is diagonalizable over $\mathbb{R}$ (as it should be).

Both these matrices were diagonalizable over $\mathbb{C}$; since their minimal polynomials have degree 3, necessarily their (complex) eigenvalues are all distinct. In the next example this is not the case:

Example 3. Let

$$C = \begin{vmatrix} 0 & -2 & 2 & 6 \\ 2 & 4 & -1 & -5 \\ -3 & -4 & 3 & 7 \\ 1 & 2 & -1 & -3 \end{vmatrix}.$$

To compute the minimal polynomial of C we start, as in Example 1, by applying the iterates of C to e_1. We get

$$Ce_1 = \begin{vmatrix} 0 \\ 2 \\ -3 \\ 1 \end{vmatrix}, \quad C^2e_1 = \begin{vmatrix} -4 \\ 6 \\ -10 \\ 4 \end{vmatrix}, \quad C^3e_1 = \begin{vmatrix} -8 \\ 6 \\ -14 \\ 6 \end{vmatrix}, \quad C^4e_1 = \begin{vmatrix} -4 \\ -8 \\ 0 \\ 0 \end{vmatrix}.$$

It is easy to check that $\{e_1, Ce_1, C^2e_1, C^3e_1\}$ are linearly independent and that

$$C^4e_1 - 4C^3e_1 + 8C^2e_1 - 8Ce_1 + 4e_1 = 0;$$

therefore $\deg \mu_{C,e_1} = 4$ and, as in Example 1, we can conclude that

$$\mu_C(t) = \mu_{C,e_1}(t) = t^4 - 4t^3 + 8t^2 - 8t + 4.$$

The standard sequence associated with μ_C starts with

$$p_0(t) = t^4 - 4t^3 + 8t^2 - 8t + 4, \quad p_1(t) = 4t^3 - 12t^2 + 16t - 8,$$
$$p_2(t) = -t^2 + 2t - 2.$$

Since p_2 divides p_1, it is the last polynomial in the sequence; since it is not constant, we conclude that C is not diagonalizable even over $\mathbb{C}$ (and in particular it cannot have four distinct eigenvalues).

Our final example involves a minimal polynomial of degree strictly less than the dimension of the space:

Example 4. Let

$$D = \begin{vmatrix} 2 & -2 & 2 & 8 \\ -2 & 4 & -1 & -9 \\ 1 & -4 & 3 & 11 \\ -1 & 2 & -1 & -5 \end{vmatrix}.$$

To compute the minimal polynomial of D we start again by applying the iterates of D to e_1. We get

$$De_1 = \begin{vmatrix} 2 \\ -2 \\ 1 \\ -1 \end{vmatrix}, \quad D^2e_1 = \begin{vmatrix} 2 \\ -4 \\ 2 \\ -2 \end{vmatrix} = 2De_1 - 2e_1;$$

therefore $\mu_{D,e_1}(t) = t^2 - 2t + 2$, and we cannot conclude right now that $\mu_{D,e_1} = \mu_D$. Proceeding with the computations we get

$$De_2 = \begin{vmatrix} -2 \\ 4 \\ -4 \\ 2 \end{vmatrix}, \quad D^2e_2 = \begin{vmatrix} -4 \\ 6 \\ -8 \\ 4 \end{vmatrix} = 2De_2 - 2e_2; \quad De_3 = \begin{vmatrix} 2 \\ -1 \\ 3 \\ -1 \end{vmatrix}$$

$$D^2e_3 = \begin{vmatrix} 4 \\ -2 \\ 4 \\ -2 \end{vmatrix} = 2De_3 - 2e_3; \quad De_4 = \begin{vmatrix} 8 \\ -9 \\ 11 \\ -5 \end{vmatrix}, \quad D^2e_4 = \begin{vmatrix} 16 \\ -18 \\ 22 \\ -12 \end{vmatrix} = 2De_4 - 2e_4;$$

hence we have $\mu_{D,e_2} = \mu_{D,e_3} = \mu_{D,e_4} = \mu_{D,e_1}$ and $\mu_D(t) = t^2 - 2t + 2$. In particular, D has (at most) two distinct complex eigenvalues.

The standard sequence associated with μ_D is

$$p_0(t) = t^2 - 2t + 2, \quad p_1(t) = 2t - 2, \quad p_2(t) = -1.$$

Since p_2 is constant, D is diagonalizable over $\mathbb{C}$. The number of variations in sign of the sequences $(1, -2, -1)$ and $(1, 2, -1)$ is

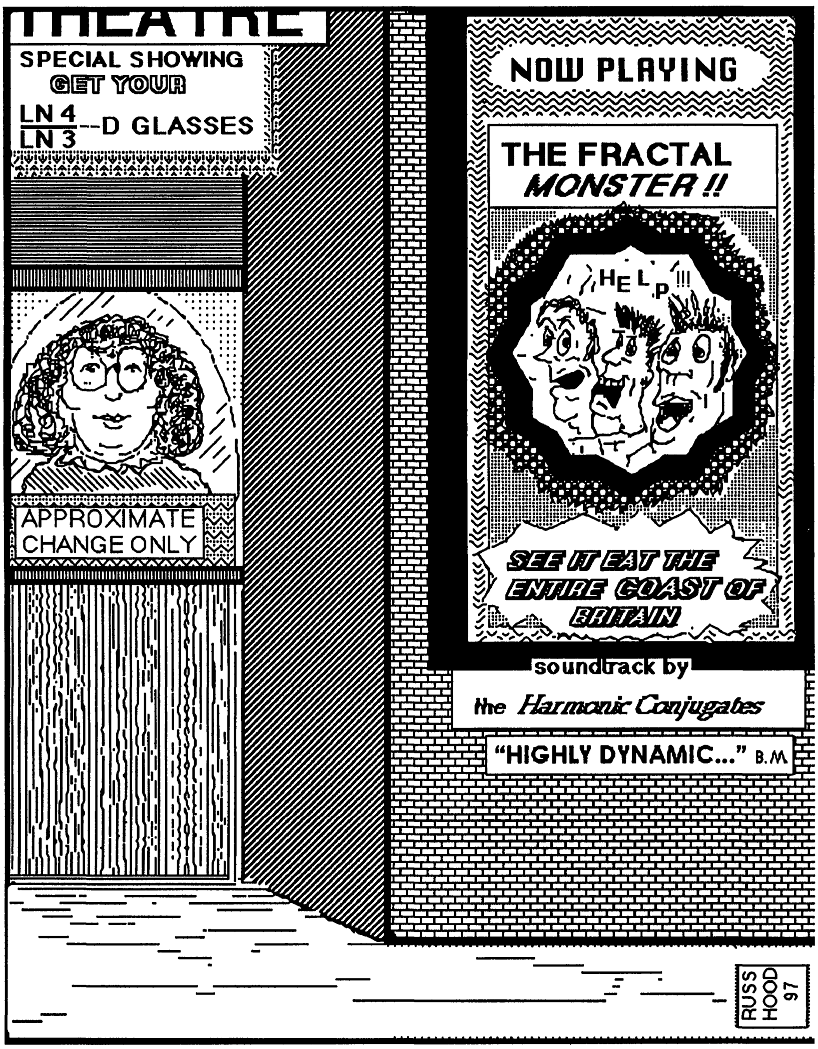
$$V_- - V_+ = 1 - 1 = 0 < 2 = \deg \mu_D,$$

and so D is not diagonalizable over $\mathbb{R}$.

REFERENCE

1. N. Jacobson, *Basic Algebra, I.* Freeman, San Francisco, 1974.

Dipartimento di Matematica
Università di Ancona
Via Brecce Bianche
60131 Ancona, Italy
abate@anvax1.unian.it



Contributed by Russell H. Hood, California State University, Sacramento

Energy Arguments in the Theory of Algorithms

Eric Bach

My goal is to convince you that physical arguments can provide valuable insight into the behavior of algorithms. In particular, several classic results from combinatorics and the theory of algorithms are easy to prove, once you have an appropriate physical model. These demonstrations have as their theme the simple but powerful idea of energy conservation: in any closed physical system, the total energy is constant.

Certainly, physical ideas have influenced the theory of algorithms in other ways. One famous example is the concept of entropy, which has found wide use in information theory. For example, you are probably familiar with the result that $\log_2 n!$ comparisons are required to sort n keys. I would like to argue here that energy is equally useful, and deserves a permanent place in your toolbox. Furthermore, abstracting the essential ideas in energy conservation arguments leads to the method of amortized analysis, for which this article can serve as an introduction.

1. PATH LENGTHS IN TREES. Let T be a binary tree, in which each node has at most 2 children. There is a unique node, called the *root*, with no parents. In the study of searching algorithms, an important role is played by the *internal path length*, defined by the following sum over the nodes of T:

$$P(T) = \sum_x [\text{number of edges between } x \text{ and the root}].$$

We also consider the *external path length*, defined by a sum over the "missing" children of the nodes of T:

$$E(T) = \sum_y [\text{number of edges between } y \text{ and the root}].$$

It is traditional to call these missing children *external nodes*.

It is well known that we minimize these functions by putting all the leaves of T on two adjacent levels, or if possible, on one level. We now show how this follows immediately from physical considerations.

We imagine that the tree grows upward from the root, in a uniform gravitational field. Thinking of each node as having unit mass, we assign the root a potential energy of 0, its children each a potential of 1, its grandchildren 2, and so on. Evidently $P(T)$ is the total potential energy of the tree, which is minimized by successively putting as many nodes as possible at each level. This gives the result for $P(T)$.

To derive the result for $E(T)$, we give the external nodes unit negative mass. (Anyone bothered by negative mass can replace the gravitational field by an electric one and use positive and negative charges.) We now imagine a disassembly process, each step of which replaces one of the highest leaves of T and its two external node children by one external node. Each disassembly step increases the

potential energy of the tree by 2. If T has n nodes, there are n steps in this process, which transforms a tree of potential energy $P(T) - E(T)$ into one with zero energy. Therefore, by conservation of energy,

$$E(T) = P(T) + 2n,$$

and any tree minimizing $P(T)$ must also minimize $E(T)$.

A corresponding relation for d-ary trees,

$$E(T) = (d-1)P(T) + dn,$$

can be proved in a similar manner.

2. OPTIMAL SEARCHING ALGORITHMS. We can apply the path length results of the preceding section to algorithms that search for a given key x among $x_1, \ldots, x_n$. These algorithms are to be composed of queries that determine whether x is larger than, equal to, or greater than another key x_i. Each algorithm determines a binary tree as follows. The root of the tree is labelled by the first key compared to x, and the left and right subtrees are determined by the queries used when x is less than or greater than the root, respectively.

Considering the root as level 1, we assign each node at level i the potential energy i. Thus, the energy of a node is the work (number of queries) needed to reach it, and if all nodes are equally likely to be sought, the average work is the total potential energy of the tree, divided by the number of nodes. This characterizes the average-case optimal search algorithms: they are exactly those whose leaves lie on two adjacent levels, or if possible, on one level. When $n = 2^k - 1$, there is only one such tree, and we recover the usual binary search algorithm.

A related problem was studied by Huffman **[5]** and can be phrased in our terms as follows. We are given particles of positive masses $p_1, \ldots, p_n$. Place these at the external nodes of a binary tree so as to minimize the total potential energy. The optimal tree does not depend on our units for mass, so we may as well assume that the p_i's sum to 1. This then corresponds to a game in which we must identify one of n possible items, occurring with the probabilities $p_1, \ldots, p_n$, using queries of the form "is $x \in S$"?

We now derive Huffman's algorithm for constructing an optimal tree, which determines a search strategy minimizing expected cost. In any minimal tree T, the two smallest masses p and q must be at the farthest distance from the root, for if not, we could swap one of them with the farthest leaf and decrease the potential. We don't change the potential by moving masses horizontally, so we may as well make them the two children of some internal node x. Now, delete these two children and give their total mass $p + q$ to the (now external) node x. This produces a new tree T'. This transformation releases an amount of energy equal to $p + q$, so the weighted external path lengths I_w must satisfy

$$I_w(T) = (p + q) + I_w(T').$$

Recursively generating T' by the same process, we solve the problem.

In deriving Huffman's algorithm, we have assumed that each node is a unit distance away from its parent. It is interesting to ask if we can further reduce the potential of T by allowing these lengths to be fractions. To prevent trivialities, we introduce the following constraint: if a node has two children at distances ℓ and r, then $2^{-\ell} + 2^{-r} = 1$. The resulting algorithm is the same as Huffman's, except that it places p and q at the distances $\log_2((q+p)/p)$ and $\log_2((q+p)/q)$ away

from their parent. You can verify that the resulting tree has potential energy

$$H(p_1,\dots,p_n) := \sum_{i=1}^{n} -p_i \log_2 p_i,$$

and that this never exceeds the potential energy of a tree constructed by Huffman's algorithm.

By this result, any binary tree with unit edge lengths and masses $p_1,\dots,p_n$ at the external nodes, summing to 1, has energy at least H. This relation is the source of many lower bounds on the performance of algorithms. For example, to prove the result on sorting mentioned in the introduction, we take (p_i) to be the uniform distribution on the $n!$ possible permutations of the input. Since H is the well-known entropy function, this provides one link between energy and information. For some others, see [**10**, p. 50].

3. COMPUTING MAX AND MIN. We now prove than any algorithm determining the minimum and maximum of a set of n keys by comparing them must use at least $3n/2 - 2$ comparisons in the worst case. This result is due to Pohl [**11**].

One way to prove results of this type is to provide some rule whereby the queries made by the algorithm can be used to construct a worst-case input for it. The algorithm is thus "hoist by its own curiosity", and the rule by which this is done is called an *adversary strategy*.

We explain a strategy sufficient to get the result, in physical terms. Each key will be in one of the four conditions V (virgin—never compared), W (won but never lost), L (lost but never won), and E (eliminated from contention). For these we have the following state diagram:

$$\begin{tikzcd} & V \arrow[dl] \arrow[dr] & \\ L \arrow[dr] & & W \arrow[dl] \\ & E & \end{tikzcd}$$

Imagine keys, represented by marbles, falling through this state diagram. We assign each marble an energy equal to its height, so the marbles in V have energy 2, marbles in W or L have energy 1, and marbles in E have energy 0. The total energy of a configuration is the sum of the energies of all the marbles. The adversary's rule is simple: whenever a query is presented, answer it in such a way as to minimize energy loss, but consistent with the answers already given.

Suppose, for example, in computing the max and min of x, y, z, the algorithm determines that $x < y$, giving a configuration of energy 4, and then compares x to z. If $x < z$, the energy becomes 3, whereas if $z < x$, it drops to 2. Since both answers are consistent, the adversary chooses the first alternative, forcing the algorithm to make an additional query.

The algorithm must get from the state in which all marbles are in V (energy $2n$) to a state in which L and W have one marble each, with the rest in E (energy 2). Each comparison drops the energy by 0, 1, or 2. Suppose there are M_i comparisons that drop the energy by i, for $0 \le i \le 2$. Then

$$M_1 + 2M_2 = 2n - 2$$

by energy conservation, whereas

$$n/2 \geq M_2,$$

because the adversary can force an energy loss at most 1, except for V-V comparisons, of which there can be at most $n/2$. Using these relations, we find that $M_1 + M_2 \geq 3n/2 - 2$, which gives a lower bound on the number of comparisons.

We also note that our model suggests an optimal algorithm. For simplicity we assume that n is even. To reduce the potential energy as quickly as possible, we first do $n/2$ V-V comparisons, then determine the maximum of W and the minimum of L. This uses

$$n/2 + (n/2 - 1) + (n/2 - 1) = 3n/2 - 2$$

key comparisons, which is best possible.

4. FINDING THE SECOND LARGEST KEY. We now prove a theorem, due to Kislitsyn [6], that any algorithm using pairwise comparisons to find the largest two elements among n keys must do at least $n + \log_2 n - 2$ comparisons on some input.

We use a physical model to study algorithms for this problem as well. We think of the n keys as particles, each of which starts off with unit energy. When two keys are compared, the result is a collision that transfers all of the energy to one of the particles, reducing the other's to zero. It is easy to see that a key has positive energy if and only if it has never lost a comparison.

By conservation of energy, the maximum key ends up with n units of energy. From this, we can see that at least $n - 1$ comparisons must be made to determine the maximum, because every other key must give up its initial energy allotment.

We now give our adversary the job of making the second largest key as difficult to find as possible. The adversary uses Matthew's rule—"them that gots shall get"—in the following way. When two keys with positive energy are compared, the winner of the comparison is always one with the largest energy. Ties and queries involving keys of zero energy may be handled in any consistent manner.

Suppose the maximum key was compared to k other keys. Then it must have had at least $n/2$ units of energy before its last comparison, at least $n/4$ before its next-to-last comparison, and so on. Since its initial energy was 1, we must have $1 \geq n/2^k$, so $k \geq \log_2 n$. We now observe that $k - 1$ of the keys compared to the maximum must lose one additional comparison, since these must also be proved not to be second largest, and the remaining $n - k - 1$ keys must lose at least one comparison. The total number of comparisons is bounded below by

$$k + (k - 1) + (n - k - 1) \geq n + \log_2 n - 2.$$

Taking into account that the number of comparisons must be integral, it can be shown that this bound is best possible.

5. PATTERN MATCHING MACHINES. We now discuss a method, due to Knuth, Morris, and Pratt [8], for finding whether a given string, called the *pattern*, occurs in a text file. Suppose we are looking for $AAAA$ in our file. Then the algorithm constructs a finite-state machine, which in this case is

$$(\text{start}) \rightleftarrows A \rightleftarrows A \rightleftarrows A \rightleftarrows A \rightarrow (\text{success})$$

Searching works as follows. If we are in the start state we read the next character of the file and advance the state to the right. If we are in a state labelled

by a character, we see if the last character read matches it. If so, we read another character and advance to the right; if not, we follow the leftward arrow to a previous state, without reading a new character. Arrival at the success state signals that the pattern has been found.

The key property of this algorithm is that at most $2n$ character comparisons are needed to search a file of length n, once the machine has been constructed. We now prove this using a physical argument. Suppose the pattern has m characters. Starting from 0, we number the states so that each successful match increases the state number by 1. The states are thus $0, 1, 2, \ldots, m + 1$. Let us agree that when the machine is in state i, it contains i units of energy. (Think, if you like, of a spring or some other gadget that gets compressed as more and more characters are matched.) Each unsuccessful comparison releases at least 1 unit of energy, because the state number goes down. On the other hand, over the entire history of the algorithm, we feed at most n units of energy into the machine, because reading a new character is the only way to put energy in. Thus, there can be at most n unsuccessful comparisons. Also, there are at most n successful comparisons, since each match causes another input character to be read. Therefore, the total number of comparisons is at most $2n$.

One can show that the machine can be constructed in the following "cannibalistic" way: search for the pattern inside itself. More precisely, if the pattern is $p_1 \cdots p_m$, we search for it in $t = p_2 \cdots p_m$. The index of the first pattern character compared to the text character p_i determines the backward link. Granting this, then, at most $2m + 2n$ character comparisons are needed to find a pattern of length m in a file of n characters.

6. ABSTRACT POTENTIAL FUNCTIONS AND EUCLID'S ALGORITHM. All of our examples have the following theme in common. We assign a number, the energy, to each possible state of a system, and use this number to draw conclusions about how it evolves over time. Even if we don't look inside the system, but only keep track of the changes in energy, we can still write the net energy change as the sum of energy changes for the various steps that are taken.

Linking these energy changes to the computational effort involved in making them has turned out to be a powerful tool in the analysis of algorithms. In the literature, this is usually called *amortized analysis* or the *potential function* method; see [**13**]. In this more abstract approach, we are no longer restricted to state functions that come from a physical model.

I close this article with a potential-function proof that the number of bit operations used by Euclid's algorithm is within a constant factor of the number used by the ordinary multiplication algorithm. This result is due to Collins [**4**].

Henceforth, all logarithms will be to the base 2. We also use $O(f)$ to indicate an unspecified function that is bounded by a positive constant times f, and $\lfloor x \rfloor$ to denote the greatest integer less than or equal to x.

We first discuss our cost model. If u is a natural number, its length in binary notation is

$$\lg u := \lfloor \log u \rfloor + 1.$$

When $u \geq 2$, we have $(\lg u)/2 \leq \log u \leq \lg u$. We assign a multiplication of u by v the cost $(\lg u)(\lg v)$, and a division expressing $u = qv + r$ with $0 \leq r < v$ the cost $(\lg q)(\lg v)$. Up to constant factors, these are the number of bit operations used by the ordinary algorithms.

We think of Euclid's algorithm as starting with the inputs $u \geq v > 0$, and replacing the pair (u, v) by $(v, u \bmod v)$, for as long as v is positive. When $(d, 0)$ is reached, d is the greatest common divisor of u and v. More details and a proof of correctness can be found in [**7**, pp. 14 ff.].

Our task is to prove that Euclid's algorithm has total cost $O((\lg u)(\lg v))$. The idea of the proof is to assign a non-negative potential to each state of the algorithm and bound each division's cost by a constant times the resulting potential change. Since we can't use more potential than we started with, the total potential change is bounded, and so is the total computation cost.

We now choose a potential function. There is a natural tradeoff involved in this choice. If we choose a simple function, it may be difficult to relate it to the running time of an individual step. A function for which this is easy to do, however, can be complicated and hard to work with. Our choice is to assign (u, v) the potential $(\log u)(\log v)$. Since the algorithm terminates with $v = 0$, we must treat the last division step specially.

Consider a division step of the algorithm, not the last one. This results in $u = qv + r$, with $0 < r < v$. The change in potential for such a step is

$$
\begin{aligned}
(\log u)(\log v) - (\log v)(\log r) &= (\log v)(\log (u/r)) \\
&> (\log v)(\log (u/v)) \qquad [\text{since } r < v] \\
&\geq (\log v)(\log q). \qquad [\text{since } q \leq u/v].
\end{aligned}
$$

Unfortunately this doesn't tell us much when $q = 1$, but there is another lower bound for this case. Since $u = v + r$ and $0 < r < v$, we have $u/r \geq v/r + 1 \geq 2$. Therefore,

$$(\log u)(\log v) - (\log v)(\log r) = (\log v)(\log (u/r)) \geq (\log v)(\log 2).$$

In every division step but the last, then, the potential is reduced by at least

$$(\log v)(\log \max \{q,2\}) \geq \frac{(\lg v)(\lg q)}{4}.$$

Now suppose the algorithm uses n division steps. Let E_i be the potential after the i-th division step. Since all steps but the last have a positive remainder, the cost of the first $n - 1$ steps is

$$\sum_{i=1}^{n-1} (\lg q_i)(\lg v_i) \leq 4 \sum_{i=1}^{n-1} (E_{i-1} - E_i) \leq 4E_0 \leq 4(\lg u)(\lg v).$$

The last division costs no more than $(\lg u)(\lg v)$, so the result is proved.

By considering how the potential function $\log u$ changes after two division steps, it can be shown that the number of division steps in Euclid's algorithm is $O(\lg u)$. This result is usually ascribed to Lamé, although he was not the first to prove it; see [**12**].

7. FURTHER READING. In considering physical models, my original goal was to find intuitive ways to verify results that are often proved by calculation. The reader may wish to compare the proofs given here with the following treatments in the literature. Sedgewick [**14**, p. 39] related the path lengths of trees by examining how they change as a binary tree is built. Knuth [**7**, p. 405] gave the corresponding relation for d-ary trees, without proof. Manber [**9**, p. 147] emphasized inductive design of algorithms, using Huffman's construction as an example. Baase [**1**, pp. 126–133] gave the adversary strategies used in Sections 3 and 4. Cormen, Leiser-

son, and Rivest [**3**, p. 872], following a suggestion of Tarjan [**13**], used amortization to study pattern-matching automata. A potential-function proof of Collins's result appears in [**2**], but this relied on Lamé's theorem.

REFERENCES

1. S. Baase, *Computer Algorithms* (Second Edition), Addison-Wesley, Reading, 1988.
2. E. Bach, Algorithmic number theory. Technical Report, ICR Distinguished Visitor Series, Computer Science Dept., University of Waterloo, Waterloo, Ontario, 1991.
3. T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, MIT Press, Cambridge, 1990.
4. G. E. Collins, Computing time analyses for some arithmetic and algebraic algorithms, *Proc.* 1968 *Summer Institute on Symbolic Mathematical Computations*, IBM Federal Systems Center, 1968, pp.195–231. Also Technical Report 36, Computer Sciences Dept., University of Wisconsin, Madison, Wisconsin, 1968.
5. D. A. Huffman, A method for the construction of minimum-redundancy codes, *Proc. IRE* 40 (1952) 1098–1101.
6. S. S. Kislitsyn, On the selection of the kth element of an ordered set by pairwise comparisons, *Sibirskii Mat. Z.* 5 (1964), 557–564. In Russian. English summary in Math. Reviews 29, #2198; Corrigendum in *ibid.*, p. 1437.
7. D. E. Knuth, *The Art of Computer Programming*: *Volume* 1, *Fundamental Algorithms* (Second Edition), Addison-Wesley, Reading, 1973.
8. D. E. Knuth, J. H. Morris, Jr., and V. R. Pratt, Fast pattern matching in strings, *SIAM J. Comput.* 6 (1977), 323–350.
9. U. Manber, *Introduction to Algorithms*: *A Creative Approach*, Addison-Wesley, Reading, 1989.
10. A. Rényi, *A Diary on Information Theory*, Wiley, Chichester, 1987.
11. I. Pohl, A sorting problem and its complexity, *Comm. ACM* 15 (1972), 462–464.
12. J. O. Shallit, Origins of the analysis of the Euclidean algorithm, *Historia Math.* 21 (1994), 410–419.
13. R. E. Tarjan, Amortized computational complexity, *SIAM J. Alg. Disc. Methods* 6 (1985), 306–318.
14. R. Sedgewick, *Algorithms* (Second Edition), Addison-Wesley, Reading, 1988.

Computer Sciences Department
University of Wisconsin
Madison, WI 53706
bach@cs.wisc.edu

News items for the MONTHLY twenty-five years ago...

Professor J. C. Eaves, Chairman of the Allegheny Mountain Section, reported that for some time the Section had been concerned with the sequence of events at its annual meeting. The Section meeting began with papers of interest to a few and ended with more of these, so that many did not come early, and too many left at noon, avoiding both the business meeting and the afternoon papers session.

...p. 160, vol. 79, 1972

Quine's NF—60 Years On

Thomas Forster

Sixty years ago in this MONTHLY, the distinguished American philosopher W.V. Quine published a novel approach to set theory. The title was "New Foundations for Mathematical Logic" [**6**]. The diamond anniversary is being commemorated by a workshop in Cambridge (England) and comes at a time of rapid increase of interest in the alternatives to the hitherto customary *Zermelo-Fränkel* set theory, which promises a new lease of life for the axiomatic system now known as '*NF*'; its creator remains in good health too. Although he is best known to a wider public for his philosophical writings, his most enduring and most concrete legacy for the next fifty years may well turn out to be his most mathematical: he gave us *NF*.

Set theory is the study of sets, which are the simplest of all mathematical entities. Let us illustrate by constrasting sets with groups. Two distinct groups can have the same elements and yet be told apart by the way those elements are related. Sets are distinguished from all other mathematical fauna by the fact that a set is constituted solely by its members: two sets with the same members are the same set. To use a bit of jargon from another age, sets are properties *in extension*. As a result, all set theories have the axiom of extensionality: $(\forall xy)(x = y \leftrightarrow (\forall z)(z \in x \leftrightarrow z \in y))$: they differ in their views on which properties have extensions.

Since set theory first sprang on the scene about a hundred years ago there has been a tendency to attempt to use this simplicity to simplify and illuminate the rest of mathematics by translating (perhaps a better word is *implementing*) it into set theory. After all, if we can represent all of mathematics as facts about these delightfully simple things, some facts about mathematics might become clear that would otherwise remain obscure. This same simplicity means that set theory is always a good topic on which to try out any new mathematical idea.

Early twentieth century mathematicians used the expression "The Crisis in Foundations". This crisis had many causes and—despite the disappearance of the expression from contemporary speech—has never really been resolved. One of its many causes was the increasing formalisation of mathematics, which brought with it the realisation that the paradox of the liar could infect even mathematics itself. This appears most simply in the form of Russell's paradox, appropriately in the heart of set theory. At first blush one might think that where sets are concerned any intension has an extension: this is the axiom of naïve set existence. For any property of sets there is a set containing precisely the sets with that property, all of those and no others. This leads rapidly to Russell's paradox, the paradox of the class of all sets that are not members of themselves. This is the Russell class. Is it a member of itself? Well, if it is it isn't and if it isn't it is. This is Russell's paradox. The *aperçu* that leapt to mind was that the problem is something to do with the possibility of sets being members of themselves, or to do with defining sets in terms of membership in themselves. Although these two might sound like two formulations of the same insight, they nevertheless lead to radically different resolutions, and to two traditions in set theory represented by Zermelo-Fränkel set theory

(often just called "set theory" by its votaries, and in any case universally abbreviated to 'ZF') and Quine's *NF*, which is our primary concern here.

According to the first view, the source of the trouble manifested in Russell's paradox is thinking of sets as things that even *might* be members of themselves. This critique gives rise to a conception of set (usually called the *cumulative hierarchy* conception) that is very easy to explain to people in a modern computer science culture: it is simply the idea that sets form a recursive datatype:

> The empty set is a set; any collection of sets forms a set; nothing else forms a set.

This declaration carries with it a kind of induction principle, as recursive datatype declarations always do. If we have an assertion that is true of the empty set, and is true of any set x as long as it is true of all x's members, then it is true of all sets. This induction principle is $\in$*-induction* and is a theorem scheme of ZF. It has various consequences, of which one of the easiest to show is that no set is a member of itself. Clearly the empty set is not a member of itself. If no member of x is self-membered, then x cannot be self-membered either, otherwise x would be a self-membered member of x, contradicting the assumption that there aren't any. How does this way of conceiving sets help with Russell's paradox? Since no set is a member of itself, the collection of sets that aren't members of themselves would have to be the collection of all sets, and there can't be such a thing, since it would be a member of itself, and we've just used $\in$-induction to show that no set can be a member of itself.

If one had more space it would be natural to expand at this point on how the conception of sets as a recursive datatype gives rise to all (well, almost all!) the axioms of ZF by using $\in$-induction to show that the recursive datatype is closed under operations corresponding to those axioms. However, here the only reason for discussing ZF is to explain the difference between the conception of set that underlies it and the conception of set that underlies *NF*.

The *NF* conception of sets does not identify the problem behind Russell's paradox as a problem about the kind of set we are going to allow to exist, and therefore not as one that can be solved by banishing sets that do not belong to a nice recursive datatype. It locates the problem instead in *the way the sets are defined.* It does this by appeal to a concept of *type*, very closely related to the concept of type in modern typed programming languages such as ML. In an ML program, it must be possible to assign every variable a consistent type, subject to various typing rules; the same idea occurs in *NF*. Just as in ML, where one assigns types to variables in the context of a whole program, in *NF* one gives types to variables in a *formula*, and does not give a variable a type for life. In *NF* the types are natural numbers, and if the variable 'x' in a formula ϕ is given the type n and the subformula '$x \in y$' appears in ϕ, then we must give 'y' the type $n + 1$. If '$x = y$' appears in ϕ then 'x' and 'y' must be given the same type. A formula is *stratified* if there is an assignment of types to variables that meets these constraints; otherwise it is *unstratified.* *NF*'s axioms are now very simply stated: (i) Extensionality; (ii) a scheme that says that the extension of a stratified formula is a set.

Let's try this on $\neg(x \in x)$. Clearly we will end up trying to give 'x' two distinct types and concluding that the formula is untyped. Therefore there is no axiom of *NF* saying that the collection of all sets that are not members of themselves is a

set, and so, *prima facie*, no paradox. The other paradoxes are all held at bay in the same way. I am careful not to say that they are *avoided*, for it is an open question whether or not NF is consistent, but they are all held at bay in the sense that the obvious derivation for each paradox relies on a set-existence axiom that is not available in NF because the relevant formula is not stratified.

So far so good: stratification seems to prevent the usual paradoxes from being derivable, but are there any deep reasons why one would expect it to have this effect, or is it just a happy—and perhaps merely temporary—coincidence? Naturally people have tried to find reasons why stratification ought to work in this way, and it turns out that stratification is not a purely syntactical notion. To explain why, we need a device first used by Bernays and Rieger to prove the independence of the axiom of foundation from ZF. A model $\mathscr{M}$ of set theory is a class with a binary relation on it, typically written $\langle M, \in \rangle$. Now let π be a permutation of M, and associate with M a new relation, which holds between x and y precisely if $x \in \pi(y)$. If there is a universal set in the model $\langle M, \in \rangle$ then there is one in the new structure too, because if V was the universal set of $\langle M. \in \rangle$ then $\pi^{-1}(V)$ will be the universal set under the new dispensation. The assertion that there is a universal set is stratified, and it turns out that not only is the assertion that there is a universal set preserved by such redefinitions of the membership relation by permutations, but also *every* stratified assertion is thus preserved. (Subject to some small print the converse is true too: every sentence thus preserved is equivalent to a stratified formula.) Although this equivalence tells us that the apparently purely syntactical concept of stratification does have some semantical significance, it doesn't seem to tell us that this significance has anything to do with the avoidance of paradox. The clearest manifestation of this gap in our understanding is that our insight about the meaning of stratification has not yet given rise to a consistency proof for NF.

The feeling among modern NFistes is that this fact about stratified formulæ (which I like to think of as a *completeness theorem* since it identifies a semantical and a syntactic property) is nevertheless something that should be taken seriously. The argument runs like this: I said just now that a model of set theory is a set (M, say) with a binary relation (R, say) on it. For present purposes we want to think of a model of set theory as a set M of atoms (things with no internal structure) associated with an injective map $i{:}M \hookrightarrow \mathscr{P}(M)$, from M into the power set of M, so that the original R associated with M can be recovered as the relation $a \in i(b)$ (where '$\in$' is the membership relation of the real world in which we who are contemplating the model reside). We can think of i as a coding function: each $a \in A$ "codes" a subset of A, namely $\{x \in A{:}x \in i(a)\}$. We know from Cantor's theorem (every set is smaller than its power set) that not every subset of A can be coded by a member of A, so in constructing a model of set theory we have to leave some sets of atoms uncoded by atoms. A decision on what injection i to associate with A is (among other things) a decision about which collections of atoms are to be sets. Now revisit the idea of the "permutation models" of the preceding paragraph. If π is again a permutation of A then we can define a to be a member of b not if (as at the start of this paragraph) $a \in i(b)$ but instead if $a \in i(\pi(b))$, and we obtain another model of set theory. What is the difference between these two models? Well (since i and $i \circ \pi$ have the same range) they have made the same decision about which classes of atoms are to be sets, but different views on how that decision is to be implemented: the same collections of atoms are to be sets of the model, it is just that they are not necessarily going to be coded by the same

elements of A as before. Accordingly the general feeling among *NF*istes is that stratification is the syntactical arm of a gang of concepts to do with what computer scientists call *implementation-invariance*.

But this is all very unhistorical. Let us go back to the years following 1937. *NF* was born in interesting times, and the West had other things on its mind during *NF*'s youth. The first really interesting development did not take place until 1953, when E.P. Specker in Zürich showed that *NF* refuted the axiom of choice and thereby proved the axiom of infinity [7]. This result was a most mysterious and disquieting one, best approached in the context of another result of Specker's, nine years later, that is in many ways more illuminating.

Specker's 1962 paper [9] connects *NF* with Russellian type theory in a way that neatly turns back the clock about 50 years. The syntax of Russell's type theory is very nasty, but the elements needed to tell its story can be recounted relatively easily. In Russell's type theory, as simplified by Ramsey, every set belongs to a *type*. The bottom type is a type of atoms, and thereafter type $n + 1$ consists of sets of things of type n. Every variable of the theory is constrained to range over one level only. Accordingly no allegation that the collection of all sets that aren't members of themselves is a set can even be *formulated* in this sort of theory, let alone proved. That fact was the attraction; there are of course drawbacks as well. One is that we thereby chuck out the baby with the bathwater, in the sense that as well as rendering unsayable things like the existence of the Russell class we also make certain apparently entirely innocent things unsayable as well. A specific consequence is that the Russell-Ramsey theory makes all sorts of assertions that look very similar but are actually distinct, even though in some sense one feels that they ought not to be. For example (according to Russellian type theory) there is no single empty set but an empty set at each type. The language does not enable us to say anything like $(\exists x)(\forall y)(y \notin x)$. But it can say $(\exists x_1)(\forall y_0)(y_0 \notin x_1)$, $(\exists x_2)(\forall y_1)(y_1 \notin x_2)$, $(\exists x_3)(\forall y_2)(y_2 \notin x_3)\ldots$ and so on, where the subscripts are type subscripts. The language clearly has an endomorphism executed as follows: take a formula, increase all the type subscripts in it by 1. The result is a new formula, written 'ϕ^+' if the first formula was 'ϕ'. What is the relation between ϕ and ϕ^+? In [8] Specker drew a parallel with projective geometry, which also has an automorphism like this. By interchanging 'point' and 'line', and interchanging 'lie on' with 'meet at' one can transform an assertion ϕ of projective geometry into another assertion of projective geometry, which is standardly called the *dual* of the first, and is written $\hat{\phi}$. It is standard that the dual of an axiom of projective geometry is another axiom. By induction on proofs one shows that the dual of a theorem is a theorem. But is $\hat{\phi} \leftrightarrow \phi$ a theorem? It is not obvious one way or the other. In the case of projective geometry the story has a neat solution and a happy ending (the scheme $\phi \leftrightarrow \hat{\phi}$ is equivalent to Desargues' theorem), but in the type theory case it is more interesting, and not just because now the '+' operation is not an involution. It is certainly the case that ϕ^+ is an axiom whenever ϕ is, and ϕ^+ is a theorem whenever ϕ is, but is $\phi \leftrightarrow \phi^+$ always a theorem? The example of the infinitely many statements saying that there is an empty set at each type is one that suggests very strongly that $\phi \leftrightarrow \phi^+$ ought to be a theorem!

It turns out that the scheme $\phi \leftrightarrow \phi^+$ is *not* a theorem of Russellian type theory but that it is consistent with Russellian type theory if and only if *NF* is consistent: this is Specker's 1962 theorem. This is very fitting when one reminds oneself of Quine's thinking behind the set existence axiom of NF. Quine's view—expressed in this MONTHLY 60 years ago—was that the type discipline that banished the

paradoxes from type theory did so by making it impossible to formulate certain set existence axioms (like that giving the Russell class), and that making multiple copies—one at each type—of apparently perfectly nonproblematic sets like the empty set is an unwanted side effect and not part of the solution. If we can avoid some of this duplication by means of judicious polymorphism then this is all to the good. The result was that Quine kept the type distinctions but instead of enforcing them at the level of syntax (so that '$x \in x$' would be illformed, as in Russellian type theory) enforced them merely at the stage of axioms of set existence, so that '$x \in x$' is wellformed, but its extension is not a set. A modern way to describe this development is to say that Quine obtained *NF* from Russellian type theory by relaxing its syntactic constraints by a bit of polymorphism, and that Specker's 1962 theorem makes this fact formal and explicit.

One consequence of Specker's discovery was the involvement of proof theory in *NF* studies. Any proof in *NF* of a stratified formula corresponds to a proof of a version of that formula (with type subscripts glued on) in Russellian type theory with a scheme of polymorphism: "from $\vdash \phi$ deduce $\vdash \phi^+$ and *vice versa*". This interchangeability relates the proof theory of *NF* to the proof theory of type theory and thereby places *NF* studies firmly in the mainstream of modern theoretical computer science. Once *NF* has been placed in such a context, it is natural to think about what happens to the ideas that gave rise to its birth if they are approached constructively. It is then natural in turn to see if the strange derivation of the axiom of infinity works from a constructive standpoint. It turns out that there is a sensible constructive version of *NF* in which we can prove that it is not the case that every set is finite, but (since constructively $\neg \forall xp$ is not the same as $\exists x \neg p$) we cannot—apparently—prove that there is an infinite set. When working with classical logic we are of course not hampered in this way, and if we can show that not every set is finite then V, the universe, is certainly infinite. Now according to *NF* V is a set (it is the extension of the expression '$x = x$' which is certainly stratified) and so too is its quotient under the equivalence relation "is the same size as". This quotient will also be infinite, and it will give us an implementation of the natural numbers. The contrast between the classical case and the constructive case, where although we can prove that not every set is finite, there doesn't appear to be any one set whose infinitude can be proved (and so we apparently cannot obtain an implementation of the natural numbers), suggests that it may be possible to prove the consistency of constructive *NF* by much simpler methods than will be needed to prove the consistency of *NF* itself.

There are other subsystems of *NF* for which we can in fact do more than merely piously hope for consistency proofs. Most of these achieve their consistency by restricting the number of comprehension axioms in one way or another. For example NF_2 has axioms to say that the universe is a boolean algebra under $\subseteq$ and that $\{x\}$ is always a set; *NF*O has in addition an axiom saying that $\{y\colon x \in y\}$ is a set. (The operation sending x to $\{y\colon x \in y\}$ enables us to show by induction on ϕ that $\{x\colon \phi(x, y_1 \dots y_n)\}$ is a set as long as ϕ is stratified and quantifier-free, and it is actually an $\in$-isomorphism!) NF_3 allows $\{x:\phi\}$ as long as the corresponding set existence axiom can be stratified with no more than 3 types. There is also a pair of theories arising from a *third* version of the circularity critique: perhaps it is necessary not only to create sets in order (as we do in the cumulative hierarchy conception) so that each set consists only of sets created earlier, but also to restrict the ways in which we specify sets so that we can form $\{x:\phi\}$ only if ϕ not only does not hold of things created later, but does not even *quantify over* sets created later.

The idea is that we should be allowed to form $\{x : \phi\}$ only if checking that x has the property ϕ does not involve examining sets we have not yet created. Set existence axioms obeying such a constraint are said to be *predicative* and it has been known for a long time that adding predicativity constraints makes consistency much easier to prove.

But the most interesting subsystem of *NF* doesn't arise in this way and was totally unexpected. This was *NF*U, uncovered by R.B. Jensen in 1969. If one weakens the extensionality axiom that is so central to set theory to allow for distinct empty sets ('U' for "Urelemente", which is what set theorists call empty sets: they are certainly very hard to tell apart!) but retains it for nonempty sets one obtains the system *NF*U. The corresponding manæuvre in ZF results in a system that is equiconsistent with ZF and was—before the development of forcing by Cohen in the 60's—used for independence proofs for the axiom of choice and the like. When we weaken *NF* to allow urelemente the effect is dramatically different: *NF*U is provably consistent and is very weak indeed, too weak to prove the axiom of infinity.

One could view the consistency of *NF*U merely as a vindication of Quine's insight that the type disciplines are enough by themselves to banish the paradoxes, even if we flirt with danger by playing with a bit of polymorphism, as does Holmes [**3**]. Although it certainly *is* such a vindication, it raises bigger questions than it answers. After all, if type disciplines are enough to put paradox to flight even when relaxed with polymorphism, why is there this dramatic difference in strength between *NF* with and without atoms? Clearly there is something else going on. (There is even the ghastly and largely unspoken possibility that the consistency of *NF*U might have nothing to do with stratification at all, but is purely the result of weakening extensionality (and thereby betraying set theory) and that even though *NF*U is consistent, *NF* itself isn't.)

But even if we do not yet understand clearly why *NF*U is so much weaker than *NF*, we can at least start to put this new system to use [**4**]. There is for the moment a great interest in alternatives to ZF, driven by the feeling that certain structures with non-wellfounded relations on them ought to be represented by sets. (A relation R on a set x is wellfounded if and only if for every nonempty subset $X' \subseteq X$ $(\exists y \in X')(\forall x \in X')(\neg(R(x, y)))$.) For a long time the standard implementation of ordinal numbers in ZF has been one that arranges for the (wellfounded) relation $<$ between ordinal numbers to be implemented by $\in$, and the idea is abroad that *all* binary relations between mathematical objects of interest should be thus representable by $\in$ between the sets chosen to implement those mathematical objects. Under the recursive datatype conception of sets (as in ZF) we can prove easily that $\in$ is a wellfounded relation on the universe of all sets. Consequently there is no possibility of representing the kind of illfounded relations that appear in computer science as relations between sets of ZF.

What is a suitable framework for this? A fashionable candidate about which a lot has been written recently is ZF with "antifoundation" axioms, of which a racy and entertaining treatment can be found in the recently published book [**1**]. Antifoundation axioms ensure that all binary relations between mathematical objects of interest are representable by $\in$ between the sets chosen to implement those mathematical objects. In a way this is a very unidiomatic thing to do to ZF. As we noted earlier, the recursive datatype conception of sets entails that $\in$ is a wellfounded relation. It is surely perverse to develop an axiomatic set theory on the basis of one conception of set, and then throw away that conception by

adopting axioms that are incompatible with it—thereby rendering suspect all the axioms it gave rise to. If we are to postulate sets that are forbidden by the recursive-datatype conception, then there is no point in looking to axioms arising from that conception to tell us how those sets are going to behave. Surely it makes more sense to have axioms of set existence that never owed anything to that conception in the first place. Such a set of axioms is to be found in *NF*U.

Can *NF*U in addition provide a set theoretic framework containing $\in$ -copies of all the structures we can describe, as postulated by the antifoundation axioms? It turns out that for various technical reasons antifoundation axioms are not consistent with *NF*U as they stand. They need to be restricted to hereditarily *small* sets. (A set is hereditarily small if and only if it is a small set of hereditarily small sets.) What is a small set? Fortunately there is an *embarras de richesse* of direct concepts of smallness: we could say that x is small if and only if x is wellordered, or if x is the same size as a wellfounded set, or x cannot be mapped onto the universal set, or is smaller than its power set. These last two seem a bit odd, but are actually quite natural in the context of *NF*U. According to *NF*U the universe is a set. Therefore Cantor's theorem, which says that every set is smaller than its power set, must fail. But it succeeds for some sets, and these typically tend to be smaller than those for which it fails. A slightly smoother notion is *strongly cantorian.* A set x is strongly cantorian if and only if the restriction of the singleton function to x is a set. Theorems of Jensen [**5**] and Holmes [**3**] tell us that the hereditarily strongly cantorian sets can be almost any ZF-style model we want. A place to look for substructures of models of *NF*U in which every set is small and antifoundation axioms are true would perhaps be the *greatest* fixed point for the operation $x \mapsto$ the set of small subsets of x. The least fixed point consists entirely of wellfounded sets and satisfies foundation rather than antifoundation.

There is no space in a brief retrospective like this to give adequate pointers to all the relevant literature, and I am uncomfortably aware that the work of my *Doktorvater* Maurice Boffa, the unofficial head of the Belgian school of *NF*istes is underrepresented in this survey, as is his collaboration with Marcel Crabbé and his rôle in furthering *NF* studies by supervising André Pétry and Roland Hinnion. Nobody likes to appear to be promoting his own work unduly, but sadly it really is true that the only book-length treatment of *NF* is [**2**]. This book also contains treatments of permutation models and all the subsystems of *NF* mentioned in this article. Fortunately for readers who have access to the web there is also Randall Holmes' *NF* website at `http://math.idbsu.edu/faculty/holmes.html`, which contains an exhaustive bibliography, links to other workers on *NF*, and Holmes' introduction to *NF*U.

REFERENCES

1. Barwise, K.J.J. and Moss, L. *Vicious Circles*. Cambridge University Press, 1996.
2. Forster, T.E. *An Essay on Set Theory with a Universal set*. second edition, Oxford Logic Guides, Oxford University Press, 1995.
3. Holmes, M.R. The set theoretical program of Quine succeeded (but nobody noticed). *Modern Logic* **4** (1994) 1–47.
4. Holmes, M.R. Elementary set theory with a universal set, Cahiers du centre de Logique **10** to appear.
5. Jensen, R.B. On the consistency of a slight (?) modification of Quine's NF. *Synthese* **19** (1969) 250–63.

6. Quine, W.v.O. New foundations for mathematical logic. *Amer. Math. Monthly* **44** (1937) 70–80.
7. Specker, E.P. The axiom of choice in Quine's new foundations for mathematical logic. *Proc. Nat. Acad. Sci. U.S.A.* **39** (1953) 972–5.
8. Specker, E.P. Dualität. *Dialectica* **12** (1958) 451–65.
9. Specker, E.P. Typical ambiguity. In *Logic, methodology and philosophy of science*. Ed E. Nagel, Stanford University Press, 1962.

Department of Pure Mathematics and Mathematical Statistics
16 Mill Lane
Cambridge CB2 1SB
U.K.
T.Forster@pmms.cam.ac.uk

Problems from 100 years ago in the MONTHLY...

41. *Proposed by O. W. ANTHONY, M.Sc., Professor of Mathematics, Columbian University, Washington, D.C.*

If the earth were a perfect sphere and had a frictionless surface, what would be the motion of a ball placed at a given latitude? [No solution of this problem has been received. EDITOR] p. 22

53. *Proposed by J. A. CALDERHEAD, M.Sc., Professor of Mathematics in Curry University, Pittsburgh, Pennsylvania.*

(a) What is the highest north latitude in which the sun will shine in at the north window of a building at least once in a year?

(b) How many days will it shine in at the north window of a building in latitude 41° N? p. 128

55. *Proposed by G. B. M. ZERR, A.M., Texarkana, Arkansas.*

It has been clear for 15 consecutive days, what is the chance of the 16th day being cloudy? p. 122

...vol. 4, 1897

Remarks on Sharkovsky's Theorem

Michał Misiurewicz

Recent publication of a paper on Sharkovsky's Theorem in this MONTHLY [8] is a good occasion for making several historical comments on this beautiful theorem.

The original paper of Sharkovsky [12] was published in Russian and has been translated into English only recently [13]. As a result, some authors citing [12] may be not fully aware of the contents of this paper. Moreover, there was a subsequent paper by Sharkovsky [14], that in some sense completed his theorem.

Consider the *Sharkovsky ordering* of the set of natural numbers:

$$3 \prec 5 \prec 7 \prec 9 \prec \cdots \prec 3 \cdot 2 \prec 5 \cdot 2 \prec 7 \cdot 2 \prec 9 \cdot 2 \prec \cdots$$
$$\prec 3 \cdot 2^2 \prec 5 \cdot 2^2 \prec 7 \cdot 2^2 \prec 9 \cdot 2^2 \prec \cdots \prec 2^3 \prec 2^2 \prec 2 \prec 1 .$$

Let I be either the real line or an interval. If $f: I \to I$ is a continuous map, then a set $P = \{x_1, x_2, \dots, x_n\}$ such that $f(x_1) = x_2, f(x_2) = x_3, \dots, f(x_n) = x_1$, is called a *cycle* or a *periodic orbit*. The *period* of a cycle P is the number of its elements.

The three parts of the full Sharkovsky Theorem are:

Theorem 1. *Let f: $I \to I$ be a continuous map. If f has a cycle of period n and if n appears before k in the Sharkovsky ordering, then f has a cycle of period k.*

Theorem 2. *For every k there exists a continuous map f: $I \to I$ that has a cycle of period k, but has no cycles of period n for any n appearing before k in the Sharkovsky ordering.*

Theorem 3. *There exists a continuous map f: $I \to I$ that has a cycle of period 2^n for every n and has no cycles of any other periods.*

In most papers and books dealing with Sharkovsky's Theorem, this name is applied only to Theorem 1. However, the original statement of Sharkovsky's Theorem is stronger. It is equivalent to Theorem 1 plus the assertion that if n appears before k in the Sharkovsky ordering then there exists a continuous map f: $I \to I$ with a cycle of period k but with no cycle of period n. Moreover, the arguments given in [12] also prove Theorem 2. Theorem 3 is proved in [14]. Thus, "Sharkovsky's Theorem" properly refers to the union of all three theorems.

The first proofs of Theorem 1 were difficult to follow. I remember that when I learned of this theorem, I tried to read the proof in [12]. The idea was clear, but the details were messy. This was apparently also an impression of Štefan, who wrote another proof [15]. However, when I tried to read Štefan's proof, I also found that the idea was clear, but the details were messy. Therefore I decided to write my own proof. When I tried to read it several months later, I realized that I did no better: the idea was clear, but the details were messy. The standard proof is now easy to follow complete detail; it was discovered almost simultaneously by many mathematicians (see e.g. [3], [4], [10], [16]).

The standard proof of Theorem 2 uses examples of maps having only cycles of odd period n and periods following n in the Sharkovsky ordering, and the "square root" construction. Various presentations of this proof (including [8]) are only small modifications of the proof in [12]. Many of them leave details to the reader (as for instance in [7, pp. 66–68]).

There was an interesting question connected with these proofs. Suppose f has a cycle P of period n and no cycle of any period preceding n in the Sharkovsky ordering. What does P look like? This question has been answered in **[2]**, **[6]**, and **[9]**. Problems of this type led to the development of *combinatorial dynamics*.

To prove Theorem 3, one has to give an example of what is called a map *of type* 2^{∞}. Several kinds of examples are known, but the most important are the ones that are smooth and unimodal ("unimodal" means "with one interior local extremum"). For these maps, as well as for one-parameter families containing them, one observes interesting geometric structure both in the parameter space and on the interval. This observation led to the development of the so called *Feigenbaum Theory* (see **[5**, pp. 199–238**]**). A very short proof of Theorems 2 and 3 together can be given by looking at the family of truncated tent maps (trapezoidal maps); see **[1]**. However, this proof is not constructive. The real career of Sharkovsky's Theorem began with the publication of the paper *Period three implies chaos* by Li and Yorke in this MONTHLY **[11]**, although the authors did not even know about Sharkovsky's Theorem when they wrote their paper.

REFERENCES

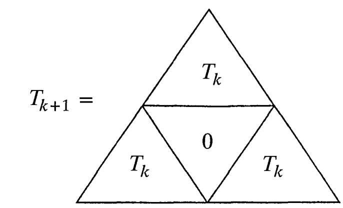
1. L. Alsedà, J. Llibre, and M. Misiurewicz, *Combinatorial Dynamics and Entropy in Dimension One*, Adv. Ser. in Nonlinear Dynamics **5**, World Scientific, Singapore, 1993.
2. L. Alsedà, J. Llibre, and R. Serra, Minimal periodic orbits for continuous maps of the interval, *Trans. Amer. Math. Soc.*, **286** (1984). 595–627.
3. L. Block, J. Guckenheimer, M. Misiurewicz, and L.-S. Young, Periodic points and topological entropy of one dimensional maps, *Global Theory of Dynamical Systems*, Lecture Notes in Math., 819, Springer, Berlin, 1980, pp. 18–34.
4. U. Burkart, Interval mapping graphs and periodic points of continuous functions, *J. Combin. Theory Ser. B*, **32** (1982), 57–68.
5. P. Collet and J.-P. Eckmann, *Iterated Maps on the Interval as Dynamical Systems*, Progr. in Phys. **1**, Birkhäuser, Boston, 1980.
6. W. A. Coppel, Šarkovskii-minimal orbits, *Math. Proc. Cambr. Phil. Soc.* **93** (1983), 397–408.
7. R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, Second Edition, Addison-Wesley, Redwood City, California, 1989.
8. S. Elaydi, On a converse of Sharkovsky's Theorem, *Amer. Math. Monthly* **103** (1996), 386–392.
9. C.-W. Ho, On the structure of the minimum orbits of periodic points for maps of the real line, preprint, Southern Illinois Univ., Edwardsville, IL (1984).
10. C.-W. Ho and C. Morris, A graph theoretic proof of Sharkovsky's theorem on the periodic points of continuous functions, *Pacific J. Math.* **96** (1981), 361–370.
11. T.-Y. Li and J. Yorke, Period three implies chaos, *Amer. Math. Monthly* **82** (1975), 985–992.
12. A. N. Sharkovsky, Co-existence of the cycles of a continuous mapping of the line into itself, *Ukrain. Math. Zh.* **16 (1)** (1964), 61–71 (Russian).
13. A. N. Sharkovsky, Coexistence of the cycles of a continuous mapping of the line into itself, *Internat. J. Bifur. Chaos Appl. Sci. Engrg.* **5** (1995), 1263–1273.
14. A. N. Sharkovsky, On cycles and structure of a continuous map, *Ukrain. Math. Zh.* **17 (3)** (1965), 104–111 (Russian).
15. P. Štefan, A theorem of Šarkovskii on the existence of periodic orbits of continuous endomorphisms of the real line, *Comm. Math. Phys.* **54** (1977), 237–248.
16. P. D. Straffin Jr., Periodic points of continuous functions, *Math. Mag.* **51** (1978), 99–105.

Department of Mathematical Sciences
Indiana University–Purdue University Indianapolis
402 N. Blackford Street
Indianapolis, IN 46202-3216
mmisiure@math.iupui.edu

Correction to: Zaphod Beeblebrox's Brain and the Fifty-Ninth Row of Pascal's Triangle

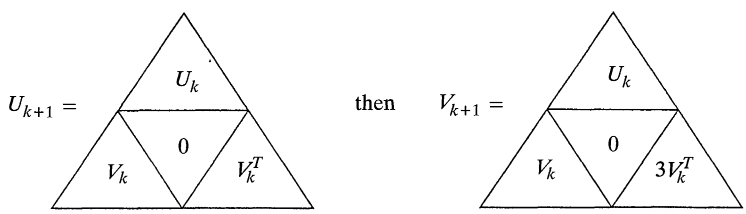
Andrew Granville

In my paper [1], we studied Pascal's Triangle modulo 2, 4, 8 and 16; and, in particular, its self-similar structure. It is well-known that the number of entries $\equiv 1 \bmod 2$ in the nth row of Pascal's Triangle is $2^{\#_2(n)}$, where $\#_2(n)$ is the number of '1's in the binary expansion of n. The proof, developed in our article, observed that if T_k denotes the top 2^k rows of Pascal's Triangle (mod 2), then



and proceeded by induction.

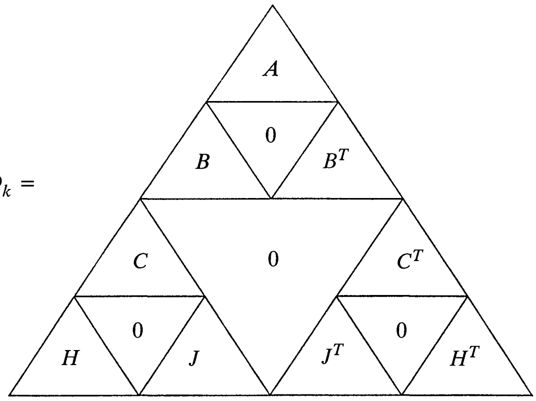
We discovered by experiment that a similar rule works with Pascal's Triangle (mod 4): if there *are not* two consecutive '1's in the binary expansion of n then there are $2^{\#_2(n)}$ entries $\equiv 1$ (mod 4), and no entries $\equiv -1$ (mod 4), in the nth row of Pascal's Triangle. On the other hand if there *are* two consecutive '1's in the binary expansion of n then there are $2^{\#_2(n)-1}$ entries $\equiv 1$ (mod 4), and $2^{\#_2(n)-1}$ entries $\equiv -1$ (mod 4), in the nth row of Pascal's Triangle. We proved this by noting that if



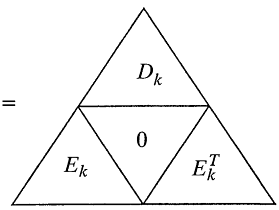
where V_k^T is the transpose of V_k, and again proceeding by induction.

The main point of [1] was that there is a version of this "self-similarity" modulo 4 (with an easily understood change of states) modulo any prime power. This also allowed us to prove that the number of entries $\equiv a \pmod{2^b}$ with a odd, $b \leq 2$ in any row of Pascal's Triangle, is either 0 or a power of 2. This is also true for $b = 3$, though the proof that we gave in [1] was faulty. Surprisingly though, $b = 4$ is an exceptional case, since exactly six entries in Row 59 of Pascal's Triangle are $\equiv 1$ (mod 16).

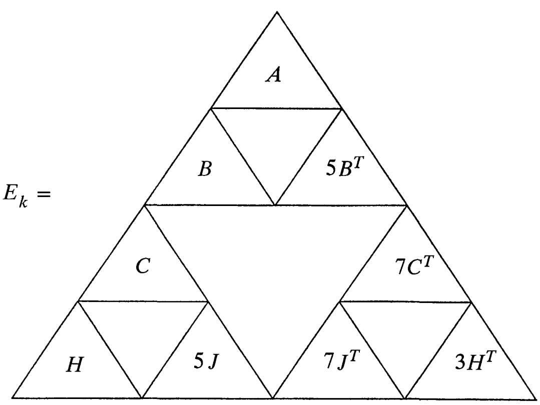
In [1], we proved that if D_k denotes the top 2^k rows of Pascal's Triangle (mod 8), with

$D_k =$ 

and if

$D_{k+1} =$ 

then

$E_k =$ 

We then attempted (see (3) in **[1]**) to write down all the cases for how the binary expansion of n determines the number of entries in row n that are $b \equiv a \pmod 8$, for $a = 1, 3, 5$ and 7; and proceeded by induction on the binary digits of n, based on the evolution of D_k into E_k. Although this method is certainly valid, Fred Howard and Ken Davis, and Jim Huard, Blair Spearman, and Ken Williams **[2]** all observed that we failed to enumerate the cases correctly; thanks to all of them for such a careful reading of my paper. Here is a correct enumeration of those cases, which follows by induction.

Let $(n)_2$ denote the binary expansion of n. If $(n)_2$ begins with "11", then we may need to "cut" the row up into four quadrants; we can discuss the first two quadrants only because of the horizontal symmetry of Pascal's Triangle. All of the statements preceding Figure 12 in **[1]** remain unchanged:

- If $(n)_2$ contains no 11 and no 101 then all odd entries are $\equiv 1 \pmod 8$
- If $(n)_2$ contains no 11 but does contain a 101 then there are an equal number of entries $\equiv 1 \pmod 8$ and $\equiv 5 \pmod 8$ and no other odd entries.
- If $(n)_2$ contains a 1111, or it contains both a 11 and a 101, then there are an equal number of entries $\equiv 1, 3, 5$ and $7 \pmod 8$, and similarly in each quadrant (when relevant).

If n does not belong to any of these cases then, in binary, it has the form

$$(n)_2 = 0\underbrace{11..1}_{t_1\ 1\text{'s}}\ \underbrace{00..0}_{u_1\ 0\text{'s}}\ \underbrace{1\ldots1}_{t_2\ 1\text{'s}}\ \ldots\ \underbrace{00..0}_{u_{m-1}\ 0\text{'s}}\ \underbrace{1\ldots1}_{t_m\ 1\text{'s}}\ \underbrace{0\ldots0}_{u_m\ 0\text{'s}}.$$

Here $1 \le t_j \le 3$ for each j, and $u_i \ge 2$ for $1 \le i \le m-1$. It is at this point that we part company from **[1]**, where we failed to distinguish certain cases that do arise. Note that we have started $(n)_2$ with a single '0', followed by the usual '1'. For example, $(825)_2 = 01100111001$.

Henceforth, we may assume that $(n)_2$ contains no 1111 nor a 101, but does contain a 11.

- If $(n)_2$ contains no 111, and n is even or $n \equiv 1 \pmod 4$, then there are an equal number of entries $\equiv 1 \pmod 8$ and $\equiv 7 \pmod 8$ and no other odd entries.
- If $(n)_2$ contains no 111 nor a 0110, and $n \equiv 3 \pmod 8$, then there are an equal number of entries $\equiv 1 \pmod 8$ and $\equiv 3 \pmod 8$ and no other odd entries.
- If $(n)_2$ contains a 111 and no 0110, and $n \not\equiv 7 \pmod 8$, then there are an equal number of entries $\equiv 1 \pmod 8$ and $\equiv 3 \pmod 8$ and no other odd entries.
- Otherwise there are an equal number of entries $1, 3, 5$ and $7 \pmod 8$.

It is possible to explain this induction proof succinctly: Let $S_n \subseteq \{1, 3, 5, 7\}$ be the set of residue classes $a \pmod 8$ such that there exists an integer j for which $\binom{n}{j} \equiv a \pmod 8$. First note that $S_{2^j} = \{1\}$ for all $j \ge 1$, $S_3 = \{1, 3\}$, and $S_{2^j-1} = \{1, 3, 5, 7\}$ for $j \geq 3$. By studying the evolution of E_k from D_k we see that

whenever $(n)_2$ contains a 101 then $5S_n = S_n$,
whenever $(n)_2$ contains a 0110 then $7S_n = S_n$,
whenever $(n)_2$ contains a 01110 then $3S_n = S_n$ and
whenever $(n)_2$ contains a 1111 then $S_n = \{1, 3, 5, 7\}$.

These observations account for the composition of S_n. If $a, b \in S_n$, then the number of entries in the nth row that are $\equiv a \pmod 8$ is equal to the number of entries that are $\equiv b \pmod 8$.

Combining these observations by studying the binary expansion of n finishes the proof.

REFERENCES

1. A. Granville, Zaphod Beeblebrox's brain and the fifty-ninth row of Pascal's Triangle, this MONTHLY 99 (1992), 318–331.
2. J. G. Huard, B. K. Spearman, and K. S. Williams, Pascal's triangle (mod 8), *European J. Combinatorics* (to appear).

Department of Mathematics
The University of Georgia
Athens, Georgia 30602-7403
andrew@sophie.math.uga.edu

It had begun to dawn on him [Frank Kennedy] that this same sweet pretty little head [Scarlett O'Hara's] was a "good head for figures." In fact, a much better one than his own and the knowledge was disquieting. He was thunderstruck to discover that she could swiftly add a long column of figures in her head when he needed a pencil and paper for more than three figures. And fractions presented no difficulties to her at all. He felt that there was something unbecoming about a woman understanding fractions and business matters and he believed that, should a woman be so unfortunate as to have such unladylike comprehension, she should pretend not to. Now he disliked talking business with her as much as he had enjoyed it before they were married. Then he had thought it all beyond her mental grasp and it had been pleasant to explain things to her. Now he saw that she understood entirely too well and he felt the usual masculine indignation at the duplicity of women. Added to it was the usual masculine disillusionment in discovering that a woman has a brain.

Margaret Mitchell, *Gone With The Wind*, Warner Books, 1994, p. 607
Contributed by Emeric Deutsch, Polytechnic University, Brooklyn, NY

A Simple Formula for π

Victor Adamchik and Stan Wagon

Dedicated to the memory of Tom Tymoczko (1943–1996), an innovative investigator into the nature of computer proofs

1. THE RADICAL BBP IDEA. In 1995 David Bailey, Peter Borwein, and Simon Plouffe [2] discovered the following shocking formula for π:

$$\pi = \sum_{k=0}^{\infty} \frac{1}{16^k}\left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6}\right).$$

This result is shocking because it can be used to generate the nth base-16 digit of π without having to look at any prior digits. And, so long as n is less than a billion or so, the entire computation can be carried out with 16-digit numbers. This is a radical idea, since all previous algorithms for the nth digit of π required the computation of all previous digits, and the use of d-digit arithmetic in the computation. For more details of the fairly easy argument that leads from the BBP formula to an algorithm for far-out hex digits of π see [1] or [2].

Proving the BBP formula is not difficult. But that misses the main point: How did they find it? In short, they had a hunch that such a formula might exist and they searched for it using high-precision approximate reals, a high-performance SGI workstation, and the PSLQ algorithm [3], [4]. In this note we show how a simpler formula of this type can be discovered in such a way that a proof accompanies the discovery. We will present only a single result. Several more formulas of this type can be found [1].

Before leaving the BBP formula, here is a proof that it is correct using *Mathematica* to perform the summation.

```
FullSimplify[TrigToExp[FullSimplify[
```

$$\sum_{k=0}^{\infty} \frac{1}{16^k}\left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6}\right)]]\,/.$$

```
a_ Log[b_] + a_ Log[c_]:>a Log[b c]].
```

π

This "proof" is of very little value, for it gives us no insight whatsoever. Some might even say that it is not truly a proof! But in principle, such a computation *can* be viewed as a proof. There are some subtleties. Some types of computations come along with certificates that allow verification; for example, if a computer churns out an indefinite integral, the result can be differentiated to see if it agrees with

the integrand. *Mathematica* does not provide such certificates for sums, but recent work of Wilf and Zeilberger has shown that certain sums, such as the ones that occur here, do carry certificates, and implementing the production and verification of such certificates has in fact been done (see **[5]**). So it is true that, provided one uses the latest work on symbolic summation algorithms, a computation such as the above can be taken to be a proof.

But such a proof is not very helpful. The real power of sophisticated symbolic software is that this first computation provides the starting point for an investigation that yields both deeper understanding of the formula and, with luck, some new formulas. That sort of investigation is what we carry out here. It turns out that the sums that arise in this note can be transformed to integrals, and then antiderivatives can be checked so that a proper standard of proof is maintained. We show how to do that at the end of Section 2, but we start our work with the reasonable working assumption that the results of `Sum` are correct.

2. DISCOVERY AND PROOF. Suppose we wish to see if π can be expressed in the following form (we examined several such forms and are presenting here the simplest one that worked).

$$\pi = \sum_{k=0}^{\infty} \frac{(-1)^k}{4^k}\left(\frac{a_1}{4k+1} + \frac{a_2}{4k+2} + \frac{a_3}{4k+3} + \frac{a_4}{4k+4}\right).$$

We just feed the general sum to *Mathematica* (we used version 3.0.0; other versions may yield slightly different forms).

$$\texttt{Simplify[FunctionExpand[}\sum_{k=0}^{\infty} \frac{(-1)^k}{4^k}\left(\frac{a_1}{4k+1} + \frac{a_2}{4k+2} + \frac{a_3}{4k+3} + \frac{a_4}{4k+4}\right)\texttt{]]}$$

$$\frac{1}{8}\texttt{(2(4(}a_2\ \texttt{ArcCot[2]} - a_4\ \texttt{Log[4]} + a_4\ \texttt{Log[5]} + a_3\texttt{(}\pi/4 + \texttt{ArcCot[3]} - \texttt{Log[25]}/4\texttt{)))} + a_1\texttt{(}\pi + 4\ \texttt{ArcCot[3]} + \texttt{Log[25]))}$$

Now we make some simplifications, the last one based on the identity $\arctan 1 + \arctan 2 + \arctan 3 = \pi$.

```
Expand[%/.{Log[25] → 2 Log[5], Log[4] → 2 Log[2],
    ArcCot[x_]:>(π/2-ArcTan[x])}/.
    ArcTan[3] → (3 π/4-ArcTan[2])]
```

$$\frac{a_2\pi}{2} + \frac{1}{2}a_1\ \texttt{ArcTan[2]} - a_2\ \texttt{ArcTan[2]} + a_3\ \texttt{ArcTan[2]} - 2\ a_4\ \texttt{Log[2]} + \frac{1}{4}a_1\ \texttt{Log[5]} - \frac{1}{2}a_3\ \texttt{Log[5]} + a_4\ \texttt{Log[5]}$$

```
Collect[%, {π, ArcTan[2], Log[5], Log[2]}]
```

$$\frac{a_2\pi}{2} + \left(\frac{a_1}{2} - a_2 + a_3\right)\texttt{ArcTan[2]} - 2a_4\texttt{Log[2]} + \left(\frac{a_1}{4} - \frac{a_3}{2} + a_4\right)\texttt{Log[5]}$$

Now we simply search for a-values that cause all but the first summand to vanish,

and the first to equal π. This is easily done by hand, but since *Mathematica* is running:

```
Solve[{a2/2 == 1, a1/2 - a2 + a3 == 0, a4 == 0, a1/4 - a3/2 + a4 == 0}]
{{a2 → 2, a1 → 2, a3 → 1, a4 → 0}}.
```

$$\mathtt{Solve}\left[\left\{\frac{\mathtt{a}_2}{2}==1,\ \frac{\mathtt{a}_1}{2}-\mathtt{a}_2+\mathtt{a}_3==0,\ \mathtt{a}_4==0,\ \frac{\mathtt{a}_1}{4}-\frac{\mathtt{a}_3}{2}+\mathtt{a}_4==0\right\}\right]$$

$$\{\{\mathtt{a}_2\to 2,\ \mathtt{a}_1\to 2,\ \mathtt{a}_3\to 1,\ \mathtt{a}_4\to 0\}\}.$$

And so we have a new formula for π:

$$\pi=\sum_{k=0}^{\infty}\frac{(-1)^k}{4^k}\left(\frac{2}{4k+1}+\frac{2}{4k+2}+\frac{1}{4k+3}\right).$$

We reiterate that the proof comes for free along with the discovery (though for a rigorous proof one might prefer to use integrals instead of sums, as we discuss in a moment). As with BBP, our formula can be used in a digit-extraction scheme in base 4. Of course, digit extraction in base 4 is fully equivalent to the base-16 case. This method of undetermined coefficients can also be used to generate the BBP formula; we leave that task to the reader who wishes to exercise a computer algebra system.

Further explorations along these lines seem to be easier if one uses integrals instead of series. This also eases the task of producing a verifiable proof. Such a transformation, focussing on the base-4 case under discussion, is carried out as follows.

1. Define $g(i)=\sum_{k=0}^{\infty}\frac{(-1)^k}{4^k}\left(\frac{1}{4k+i}\right)$.

2. $g(i)=\sum_{k=0}^{\infty}\sqrt{2^i}\int_0^{1/\sqrt{2}}(-1)^k z^{4k+i-1}\,dz$ (easy integration).

3. $g(i)=\sqrt{2^i}\int_0^{1/\sqrt{2}}\sum_{k=0}^{\infty}((-1)^k z^{4k+i-1})\,dz$ (interchange).

4. $g(i)=\sqrt{2^i}\int_0^{1/\sqrt{2}}\frac{z^{i-1}}{1+z^4}\,dz$ (geometric series).

5. Use undetermined coefficients a_i with (4) and call on either a computer or an integration expert to get a closed-form expression for $\sum_{i=1}^{4}a_i g(i)$. It will agree with the four-transcendental expression we obtained at the beginning of this section.

Of course, many other forms can be investigated in the hope of getting more formulas for π or other constants. Sadly, it seems as if these ideas may not lead to a formula that allows extraction of base-10 digits. From one point of view, the crucial miracle that makes the above formulas work is that certain arctangents that arise are rational multiples of π. J. Buhler has shown that this can happen only in situations that are essentially equivalent to the ones above; in particular, no base-10 formula relying on this particular phenomenon exists. Still, there might be other numerical miracles that could give a base-10 formula or, more generally, other kinds of formulas or techniques for rapid extraction of base-10 digits.

REFERENCES

1. V. Adamchik and S. Wagon, π: A 2000-year-old search changes direction, *Mathematica in Education and Research* 5:1 (1996) 11–19.

2. D. H. Bailey, P. Borwein, and S. Plouffe, On the rapid computation of various polylogarithmic constants, *Mathematics of Computation* (to appear).
3. D. H. Bailey and S. Plouffe, Recognizing numerical constants, *Proceedings of the Workshop on Organic Mathematics* (electronic at www.cecm.sfu.ca/organics/).
4. H. R. P. Ferguson, D. H. Bailey, and S. Arno, Analysis of PSLQ, an integer relation finding algorithm, forthcoming.
5. M. Petkovšek, H. S. Wilf, and D. Zeilberger, $A = B$, A K Peters, Wellesley, Mass., 1996.

Wolfram Research, Inc.
100 Trade Center Drive
Champaign, IL 61820
victor@wolfram.com

Macalester College
St. Paul, MN 55105
wagon@macalester.edu

Borsuk-Ulam Implies Brouwer: A Direct Construction

Francis Edward Su

1. INTRODUCTION. The Borsuk-Ulam theorem and the Brouwer fixed point theorem are well-known theorems of topology with a very similar flavor. Both are non-constructive existence results with somewhat surprising conclusions. Most topology textbooks that cover these theorems (e.g., **[4]**, **[5]**, **[6]**) do not mention the two are related—although, in fact, the Borsuk-Ulam theorem implies the Brouwer Fixed Point Theorem.

The theorems themselves are often proved using the machinery of algebraic topology or the concept of degree of a map. That one theorem implies the other can therefore be established once one understands this machinery, but this requires background. Moreover, such proofs tend to be indirect, relying on the equivalence of these existence theorems with corresponding *non-existence* theorems. For instance, Dugundji and Granas **[3]** show that the Borsuk-Ulam theorem is equivalent to the statement that no antipode-preserving, continuous map f: $S^n \to S^n$ can be homotopic to a constant map. From this one can see that the Brouwer fixed point theorem is a special case, because it can be shown equivalent to the statement that the identity map *id*: $S^n \to S^n$ (which is antipode-preserving) is not homotopic to a constant map.

However, such an indirect approach is not really necessary, and perhaps a more direct proof would give insight as to how the two theorems are related. The purpose of this note is to provide a completely elementary proof that the Borsuk-Ulam theorem implies the Brouwer theorem by a *direct* construction, in which the existence of antipodal points in one theorem yields the asserted fixed point in the other.

2. THE THEOREMS. Let S^n denote the unit n-sphere in $\mathbb{R}^{n+1}$, i.e., all points at distance one from the origin. Two points are *antipodal* if they lie opposite each other on the sphere—i.e., $\{\mathbf{x}, -\mathbf{x}\}$ for some $\mathbf{x}$.

The Borsuk-Ulam Theorem. *Let $f:S^n \to \mathbb{R}^n$ be a continuous map. There exists a pair of antipodal points on S^n that are mapped by f to the same point in $\mathbb{R}^n$.*

This theorem was conjectured by S. Ulam and proved by K. Borsuk [**1**] in 1933. In particular, it says that if $f = (f_1, f_2, \dots, f_n)$ is a set of n continuous real-valued functions on the sphere, then there must be antipodal points on which all the functions agree. For instance, one interpretation for the case $n = 2$ is that there is always a pair of antipodal points on the earth's surface with the same temperature and barometric pressure (assuming, of course, that temperature and pressure vary continuously).

Let B^n denote the unit n-ball in $\mathbb{R}^n$. A *fixed point* for a map f from a space into itself is a point $\mathbf{y}$ such that $f(\mathbf{y}) = \mathbf{y}$. The following theorem, due to L.E.J. Brouwer, is one of the most celebrated theorems in topology:

The Brouwer Fixed Point Theorem. *Every continuous map $f:B^n \to B^n$ possesses a fixed point.*

Brouwer proved the case $n = 3$ in 1909, and Hadamard followed in 1910 with a proof for all dimensions. Brouwer gave a different proof in 1912 [**2**]. See [**3**] for more historical notes and a survey of fixed point theory.

In dimension three, the Brouwer theorem is often interpreted as follows: no matter how you slosh around the coffee in a coffee cup (as long as you do it continuously), some point is always in the same position it was before the sloshing took place (although it might have moved around in the meantime). Moreover, should you try to move this point out of its original position, you will unavoidably move some *other* point back into its original position.

3. THE IDEA. As motivation we first briefly sketch a construction that shows how the Borsuk-Ulam theorem implies the Brouwer fixed point theorem.

We choose to think of B^n as $[-1,1]^n$, the "n-cube" in $\mathbb{R}^n$. Similarly, we choose to think of S^n as the boundary of the $(n+1)$-cube $[-1,1]^{n+1}$ in $\mathbb{R}^{n+1}$:

$$S^n = \{\mathbf{x} \mid \mathbf{x} = (x_1, x_2, \dots, x_{n+1}), \quad |x_i| \le 1 \text{ and } \max |x_i| = 1\}.$$

The "cubical" n-sphere is homeomorphic to the usual n-sphere via the rays from the origin. In fact, this is an antipode-preserving homeomorphism, so the Borsuk-Ulam theorem holds for maps on cubical n-spheres. We choose to work with cubical n-spheres and n-balls because constructing and describing functions on such objects is easier in rectangular coordinates.

Given f: $B^n \to B^n$, we would like to construct a map g: $S^n \to \mathbb{R}^n$ that encodes f in such a manner that the existence of Borsuk-Ulam antipodal points for g implies the existence of a Brouwer fixed point for f.

The idea is as follows: on the cubical n-sphere, the "top" and "bottom" faces of the cube are homeomorphic copies of B^n. These are separated by an "equatorial" band, consisting of the other faces. Our task is to define a continuous function g on these three regions of S^n. On the top face, we define g in such a way that a zero of g implies a fixed point for f. We then define g on the bottom face so that the image of each point there is the negative of the image of its antipode on the top face. Such map is called *antipode-preserving*—meaning $g(-\mathbf{x}) = -g(\mathbf{x})$. If we can patch-in the equatorial region with a map that is also antipode-preserving but never zero, then the Borsuk-Ulam Theorem guarantees the existence of antipodal

points that get mapped by g to the same point. Because g is antipode-preserving, these antipodal points must get mapped to zero, which by construction cannot occur in the equatorial band. A zero for g on the top or bottom face then implies a fixed point for f.

4. THE CONSTRUCTION. We seek to construct a map $g = (g_1, g_2, \dots, g_n)$: $S^n \to \mathbb{R}^n$ that is continuous and antipode-preserving, i.e., $g(-\mathbf{x}) = -g(\mathbf{x})$.

We first construct g on the "top" and "bottom" faces. Note that each face, on which some coordinate $x_k = \pm 1$, is an n-cube. When $x_{n+1} = \pm 1$ we obtain:

$$S^n_{\text{top}} = \{\mathbf{x} \mid \mathbf{x} = (x_1, x_2, \dots, x_n, 1)\} \quad \text{and} \quad S^n_{\text{bot}} = \{\mathbf{x} \mid \mathbf{x} = (x_1, x_2, \dots, x_n, -1)\},$$

which denote the top and bottom faces of the cubical S^n. See Figure 1. Let $p : \mathbb{R}^{n+1} \to \mathbb{R}^n$ be defined by $p(\mathbf{x}) = (x_1, \dots, x_n)$, i.e., p ignores the last coordinate.

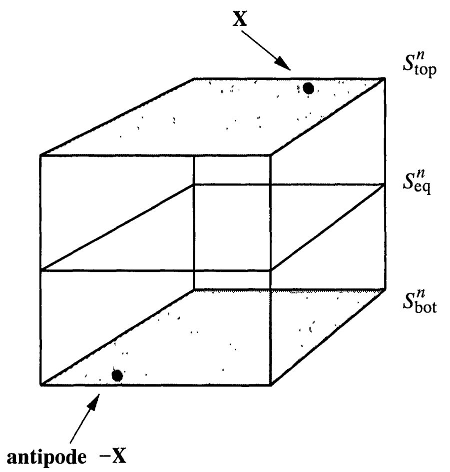


Figure 1. Top, bottom, and equator of the cubical n-sphere (here $n = 2$). An example of antipodal points is indicated.

For all $\mathbf{x}$ in S^n_{top}, define $g(\mathbf{x}) = p(\mathbf{x}) - f(p(\mathbf{x}))$. For all $\mathbf{x}$ in S^n_{bot}, define $g(\mathbf{x}) = p(\mathbf{x}) + f(-p(\mathbf{x}))$.

Since $p(-\mathbf{x}) = -p(\mathbf{x})$, one may check that $g(-\mathbf{x}) = -g(\mathbf{x})$. Thus g is, so far, antipode-preserving. It is continuous, since f and p are. If $g(\mathbf{x}) = \mathbf{0}$ then $p(\mathbf{x})$ is a fixed point for f.

Now we want to define g on the "side" faces of the cubical n-sphere so that it matches up continuously with g on S^n_{top} and S^n_{bot} and is still antipode-preserving, but is never zero on the sides. The latter is the tricky part.

One might try to extend the values of g linearly from top to bottom, but this does not guarantee that $g \neq \mathbf{0}$ on the sides. However, the following lemmas show that if we define g suitably on the equator, we can linearly extend the values from the equator to top and bottom without creating a new zero for g.

Lemma 1. *Let F be a "side" face of the cubical S^n. That is, there exists some k, $1 \le k \le n$, such that for all $\mathbf{x}$ in F, x_k is constant and equal to $+1$ or -1. Then for*

all $\mathbf{x}$ *in* $F \cap (S^n_{\text{bot}} \cup S^n_{\text{top}})$, *the coordinate function* $g_k(\mathbf{x})$ *is either* 0 *or has the same sign as* x_k.

Proof: By the definition of g, on $F \cap S^n_{\text{top}}$, $g_k(\mathbf{x}) = x_k - f_k(p(\mathbf{x}))$ and on $F \cap S^n_{\text{bot}}$, $g_k(\mathbf{x}) = x_k + f_k(-p(\mathbf{x}))$. Since f is a map to an n-cube, $|f_k| \le 1$. Hence, if $x_k = 1$, then $g_k(\mathbf{x}) \ge 0$ for all $\mathbf{x} \in F \cap (S^n_{\text{bot}} \cup S^n_{\text{top}})$. If $x_k = -1$, then $g_k(\mathbf{x}) \le 0$ for all $\mathbf{x} \in F \cap (S^n_{\text{bot}} \cup S^n_{\text{top}})$. ■

Now letS^n_{eq} denote the "equator" of S^n, i.e., $\{\mathbf{x} \in S^n \,|\, \mathbf{x} = (x_1, \dots, x_n, 0)\}$. For all $\mathbf{x} \in S^n_{eq}$, define g on S^n_{eq} by

$$g(\mathbf{x}) = p(\mathbf{x}) + \frac{g(x_1, \dots, x_n, -1) + g(x_1, \dots, x_n, 1)}{2}, \tag{1}$$

and observe that g is now antipode-preserving on the equator.

Lemma 2. *For all* $\mathbf{x} \in S^n_{eq}$, *if* $|x_k| = 1$, *then the coordinate function* $g_k(\mathbf{x})$ *is not* 0 *and has the same sign as* x_k.

Proof: Lemma 1 shows that if $|x_k| = 1$, g_k has the same sign as x_k on the top and bottom faces if it is not zero. Therefore, using (1) and $p_k(\mathbf{x}) = x_k = \pm 1$, we see that $g_k(\mathbf{x})$ is non-zero on the equator and has the same sign as x_k.

To define g on the equator we have "averaged" the values of the corresponding points on S^n_{top} and S^n_{bot} and then "lifted" that average by $p(\mathbf{x})$ (which equals x_k in

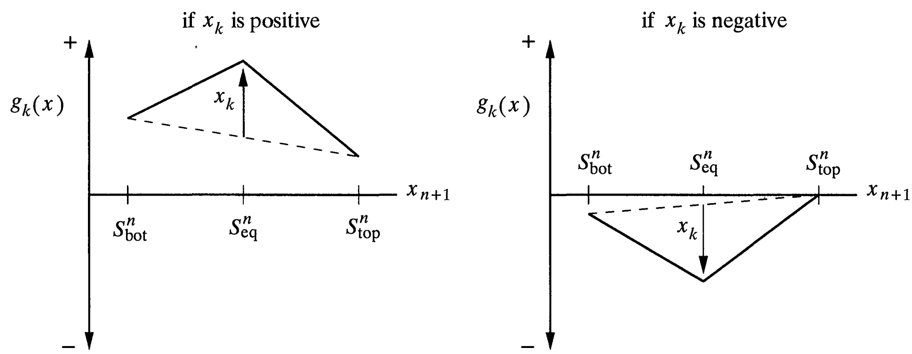


Figure 2. Extending the values of g_k linearly from the equator to top and bottom. Graphs show a cross-section of g_k values along on "longitude" of the cubical n-sphere on the faces determined by $x_k = +1, -1$.

the k-th coordinate) to pull it away from possibly being zero. See Figure 2.

We now define g continuously on the rest of S^n by extending it linearly from the equator to the values on S^n_{bot} and S^n_{top}. That is, for $0 \le x_{n+1} \le 1$, let

$$g(\mathbf{x}) = x_{n+1} g(x_1, \dots, x_n, 1) + (1 - x_{n+1}) g(x_1, \dots, x_n, 0). \tag{2}$$

For $-1 \le x_{n+1} \le 0$, let

$$g(\mathbf{x}) = -x_{n+1} g(x_1, \dots, x_n, -1) + (1 + x_{n+1}) g(x_1, \dots, x_n, 0). \tag{3}$$

Refer to Figure 2 again. Note that g is continuous and antipode-preserving.

Furthermore, it can achieve $\mathbf{0}$ only on S^n_{top} or S^n_{bot}, because of

Lemma 3. *If* $|x_{n+1}| < 1$, *then* $g(\mathbf{x}) \neq \mathbf{0}$.

Proof: Since $|x_{n+1}| < 1$, we are on a side face and therefore there exists some k, $1 \le k \le n$, for which $x_k = \pm 1$. We shall show that $g(\mathbf{x})$ cannot be zero by showing that the coordinate function $g_k(\mathbf{x})$ is non-zero.

Consider (2) and (3). By Lemmas 1 and 2, $g_k(x_1, x_2, \ldots, x_n, \pm 1)$ and $g_k(x_1, x_2, \ldots, x_n, 0)$ have the same sign as x_k, and the latter is non-zero. Moreover, $(1 - x_{n+1})$ and $(1 + x_{n+1})$ are strictly positive because $|x_{n+1}| < 1$. Equations (2) and (3) now imply that $g_k(\mathbf{x})$ is non-zero and, in fact, has the same sign as x_k. ■

Now that g is defined everywhere on S^n, the Borsuk-Ulam Theorem implies that there exists a pair $\{\mathbf{x}, -\mathbf{x}\}$ such that $g(\mathbf{x}) = g(-\mathbf{x})$. But $g(\mathbf{x}) = -g(-\mathbf{x})$, since g is antipode-preserving. Therefore $g(\mathbf{x}) = g(-\mathbf{x}) = \mathbf{0}$ which, by Lemma 3, implies that one of the pair $\{\mathbf{x}, -\mathbf{x}\}$ is in S^n_{top}. Without loss of generality, suppose it is $\mathbf{x}$. Then $g(\mathbf{x}) = p(\mathbf{x}) - f(p(\mathbf{x})) = \mathbf{0}$ on S^n_{top} implies that for $y = p(\mathbf{x}) \in B^n$, we have $f(\mathbf{y}) = \mathbf{y}$, which proves the Brouwer Fixed Point Theorem.

ACKNOWLEDGMENTS. I thank Michael Starbird for many lively discussions and Courtney Coleman for helpful stylistic suggestions.

REFERENCES

1. K. Borsuk, Drei Sätze über die n-dimensionale euklidische Sphäre, *Fund. Math.*, **20** (1933), 177–190.
2. L.E.J. Brouwer, Über Abbildung von Mannigfaltigkeiten, *Math. Ann.*, **71** (1912), 97–115.
3. J. Dugundji and A. Granas, *Fixed Point Theory*, Polish Scientific Publishers, Warsaw, 1982.
4. W.S. Massey, *A Basic Course in Algebraic Topology*, Springer-Verlag, New York, 1991.
5. J. Rotman, *An Introduction to Algebraic Topology*, Springer-Verlag, New York, 1984.
6. E.H. Spanier, *Algebraic Topology*, Springer-Verlag, New York, 1966.

Department of Mathematics
Harvey Mudd College
Claremont, CA 91711
su@math.hmc.edu

Does Mathematics Distinguish Certain Dimensions of Spaces?

Zdzisław Pogoda and Leszek M. Sokołowski

translated by Abe Shenitzer with the editorial assistance of Hardy Grant

PART I

1. DIMENSION IN PHYSICS, OR: WHAT WORLD DO WE LIVE IN? The Greeks of, say, Euclid's time talked of *diastasis*—extension—of bodies rather than of the number of their dimensions.

It was Claudius Ptolemy who advanced the idea that the extension of a body should be measured by the number of mutually perpendicular straight lines (rigid rods) issuing from a point in the body and claimed that this number cannot exceed three. Ptolemy's idea is a precursor of the idea of a system of coordinate axes used to this day in mathematics (outside topology) and in physics. His claim—we call it Ptolemy's law—is deeply rooted in Western culture and remained unchallenged until the third decade of the 20th century. The remarkableness of this fact becomes clear if we recall that the totality of the physics of the ancients, with the exception of the Pythagorean thesis of the sphericity of the Earth and of the celestial bodies, of Archimedes' law, and of the elements of the theory of simple machines, was questioned and rejected at the beginning of the modern era. Ptolemy's law was first challenged in 1919 when Theodor Kaluza, then an instructor at Königsberg, sent a letter to Einstein in which he advanced the hypothesis of a 5-dimensional universe. Kaluza did not base his hypothesis on any experimental facts. His thinking was guided by the notion that in a 5-dimensional universe he could unify gravitation and electromagnetism.

In a sense, nothing has changed since 1919. Ptolemy's law is supported by countless experiments in all areas of physics. Moreover, there is not a single empirical fact that would suggest even indirectly the existence of additional dimensions of space. At the present time, the only reason for questioning Ptolemy's law is the hope that in a multidimensional universe all elementary interactions could be combined into one. Viewed by an observer in "our" 4-dimensional spacetime, this single interaction breaks down into gravitation, nuclear forces, and electroweak interactions. The only thing that has evolved between 1919 and today is the concrete version of the unification idea. First it was the classical 5-dimensional theory of Kaluza-Klein, then the multidimensional theory of Kaluza-Klein with arbitrary gauge fields, based largely on the conceptions of De Witt and

Witten, and, as of 1984, the hopes of the physicists have focussed on the theory of superstrings.

These theories have in common the view that physical spacetime has $d = 4 + n$ dimensions of which four are directly observable and the remaining n are not—at least not with present experimental techniques. The two key questions to be answered here are: what is the value of n, i.e., the value of the dimension d? and why does the d-dimensional universe contain a distinguished subspace of dimension four? The second question is actually made up of the following two: why does the universe not have a structure analogous to that of Minkowskian spacetime, all of whose dimensions are equivalent in a well-known sense of the term, but decomposes into an "easy-to-see" component and a "difficult-to-see" component, and what is the significance of the dimensions of these components.

We have no satisfactory answer to any of these questions. The classical Kaluza-Klein theory, in which $d = 5$, was given up long ago. There is also a modern version of this theory based on supergravity. In 1981 Edward Witten, working with this version, offered quite convincing arguments in favor of the value $d = 11$. But this theory had to be abandoned three years later because of insurmountable difficulties. In its place came the theory of superstrings, which engendered great hopes and strong emotions. Some important scientists hailed it as the General Theory of Everything. By 1987 this enthusiasm decreased somewhat. Yet the theory of superstrings remains the most promising candidate for a theory that bids fair to quantize gravitation and unify all interactions. This would lead to a coherent theory of all elementary particles. Still, the theory of superstrings is only a candidate for such a theory. Also, it is, so to say, very much nonunique, in that it admits a great variety of models. The trouble is that we do not know enough to choose the "true" one of these models. Furthermore, the theory is nonunique when it comes to the dimension of spacetime. At first it seemed that if one considers bosonic strings then the theory can be internally consistent only for $d = 26$, while the theory of fermionic and supersymmetric strings, of greater interest for physicists, requires the value $d = 10$. Later, however, it was shown that what had previously been regarded as consistency was actually just simplicity: both dimensions are singled out by the theory for "pure" strings. By associating a number of physical fields with a string one can achieve a consistent string theory for any value of d. Thus the theory again gives no hint regarding the dimension of the physical world.

Suppose we accept $d = 10$ as a tentative answer to the first of our two questions. What about an answer to the second one? All we can say at this point is that we know neither the mechanism that brought into being the "easy-to-see" and the "difficult-to-see" components of the universe nor the factors that determine their respective dimensions.

In spite of the complete absence of relevant experimental data, the idea of a multidimensional universe has become a fixed part of theoretical physics and can be expected to influence the direction of its development for a long time to come. This being so, it is reasonable to accept tentatively its validity. Even if the idea turns out to be false, the question it gave rise to, namely the question of why Ptolemy's law holds, i.e., why $d = 4$, retains its relevance. In other words, a satisfactory future physical theory must not treat the dimension of physical spacetime as a free parameter whose value is determined experimentally but must be able to explain why its value is what it is.

2. DIMENSION IN MATHEMATICS: IS IT WEIGHTY? We have seen that, so far, physics provides no persuasive answer to the question of the dimension of physical space. Does mathematics? Does mathematics distinguish certain dimensions of space and are they close to 3, the only empirically distinguished dimension? More generally, is dimension a significant mathematical concept? In other words, will a mathematician agree that the question of the number of dimensions of spacetime is important or will he tell us that the number of dimensions of, say, Euclidean space conveys information that is as trivial as that conveyed by the fact that we describe the surface of the earth with respect to a rectangular coordinate system oriented in the east-west and north-south directions? If this were the viewpoint of mathematics, then we would have to admit that the empirically determined dimension of spacetime is something "local" and devoid of deeper physical significance, something akin to the empirically determined difference between the horizontal and vertical directions, something due to the local gravitational field of the earth and without a universal character. This would also imply that every physical theory in which dimension plays an important role is false in a definite sense. To illustrate what we are trying to say note that Aristotle's physics, with its division of the universe into sublunar and translunar regions, has no adequate mathematical apparatus; Euclidean geometry, known in antiquity, does not fit it in the least. We hope that these remarks make it clear that the determination of the status of the concept of dimension in mathematics is an issue of profound importance for physics.

While the mathematical notion of dimension first came up in the context of Euclidean geometry, it is actually a topological concept applicable to quite a large class of topological spaces. Some basic problems of dimension theory were solved only in the first half of the 20th century and many others remain open. Some mathematicians think that problems associated with the notion of dimension are of the utmost importance not only in topology but also in all of mathematics. There is the famous dictum of Poincaré: "I think that the most important of the theorems of *Analysis situs* [an earlier name for topology that goes back to Leibniz and was very popular in the 19th century] is the one that asserts that space has three dimensions [**1**]."

We noted earlier that the roots of dimension theory go back to the Hellenistic period. In modern mathematics the concept of dimension was initially treated intuitively and nonuniformly, namely, there were distinct definitions of dimension in geometry and in linear algebra, as well as distinct definitions of the dimension of a simplex and of the dimension of a linear subspace. Nevertheless dimension was, and continues to be, one of the most important elements of the description of space and of geometric objects. The need for a nonintuitive, rigorous definition of dimension was highlighted by remarkable discoveries made at the turn of the 20th century. Thus in 1877 Georg Cantor showed that a segment and a square have the same cardinalities, i.e., that there exists a one-to-one mapping of a segment onto a square. This implied that dimension is not a set-theoretic concept. Again, a curve can be thought of as the result of a single stroke of a pencil. Hence Camille Jordan's definition of a curve as a subset of the plane that is a continuous image of an interval. Intuition dictated the notion that such an object is 1-dimensional. But in 1890 Giuseppe Peano constructed a curve, i.e., a continuous image of an interval, that filled a square. These results showed that the dimension of a Euclidean space is not preserved either by one-to-one or by continuous mappings. There remained the open problem of whether a topological mapping, i.e., a

one-to-one bicontinuous mapping, can map a plane onto 3-space. We should add that at that time the question of dimension in topology was primarily a question of its relation to Euclidean geometry. The properties of plane figures are so different from the properties of solids that if the dimension of a Euclidean space were not a topological invariant then topology would tell us very little about its geometric properties. The problem of dimension was made more acute by the discovery of curious figures such as the Sierpiński carpet and the Menger cube (see **[2]**, **[4]**), whose dimensions cannot be guessed at a glance. It was obvious that the notion of dimension had to be reinvestigated "from the ground up."

It is probably safe to say that a breakthrough was Luitzen E.J. Brouwer's proof, provided between 1911 and 1912, of the invariance of domain (see **[2]**). This implied that if $n \neq m$ then $\mathbb{R}^n$ and $\mathbb{R}^m$ are not homeomorphic. The significance of this result is obvious: since the intuitive notion of the dimension of a Euclidean space, given by the number of coordinates, turned out to be a topological invariant, it was reasonable to talk of dimension in topology. But the nature of the new invariant remained unclear because Brouwer did not use in his proof any topological characteristics of Euclidean dimension.

During the next two decades, i.e., until the mid 1930s, mathematicians such as Henri Lebesgue, Karl Menger, Pavel Uryson, and Eduard Čech developed the foundations of a topological theory of dimension by defining dimension in a rigorous way and by proving many theorems that determined its properties; the theory applies to all classical cases (see **[5]**). Unfortunately, the theory is not unique in that there are three different definitions of the dimension of a topological space, namely the covering dimension (dim), the small inductive dimension (ind), and the large inductive dimension (Ind). These three notions coincide for metric spaces with a countable basis but diverge for more general spaces.

Fortunately, the complications that arise in the general theory do not worry the physicists because the most general spaces used in physics are differentiable manifolds. In view of Whitney's immersion theorem such a manifold can always be represented as a subset of $\mathbb{R}^k$ for a large enough k. More precisely, if the manifold has dimension n, then it is a (regularly immersed) subset of $\mathbb{R}^{2n+1}$ (see **[2]**). In the sequel we will limit ourselves to $\mathbb{R}^n$ and its subsets. This means that we can use any one of the three definitions of dimension. These are rather complicated; the simplest and the most convenient for us is the definition of small inductive dimension. We present an abbreviated version of it.

1. The dimension of the empty set is -1 ($\text{ind}\varnothing = -1$).
2. If every point of a space X has an arbitrarily small neighborhood whose boundary has dimension $\leq n-1$, then X has dimension $\leq n$ ($\text{ind}\, X \leq n$).
3. If $\text{ind}\, X \leq n$ but it is not true that $\text{ind}\, X \leq n-1$, then X has dimension n ($\text{ind}\, X = n$).
4. If $\text{ind}\, X \leq n$ is false for all n, then the dimension of X is infinite ($\text{ind}\, X = \infty$).

If we look carefully at the above definition of dimension, then we see that it includes all the intuitive perceptions that we associate with this notion. If we remove from a line a point together with a small neighborhood of this point, then the boundary of that neighborhood consists of two points (if finitely many lines intersect at the point in question, then the corresponding boundary consists of finitely many points) and so has dimension 0. Similarly, if we remove from a plane (or, more generally, from a surface) a point together with a small neighborhood of

this point, then the boundary of that neighborhood consists of a closed line and so has dimension 1. This observation can be carried over by induction to any number of dimensions.

The definition just given satisfies almost all of the requirements we are likely to associate with the notion of dimension. Its advantages are twofold. For one thing, the inductive definition of dimension of a set agrees with its "intuitive dimension." For example, the inductive dimension of "n-dimensional" figures such as a cube, a sphere, a ball, and a simplex is n (this is difficult to prove!). For another, the definition of the dimension ind enables us to assign dimension numbers to an extensive class of topological spaces that baffle our intuition. Examples of such spaces are the Cantor set (dimension 0), the Sierpiński carpet (ind $= 1$), and the Menger cube (ind $= 1$) (see [**3**]).

We used the words "almost all" in connection with the definition of ind because, upon closer inspection, it turns out to have certain flaws (we won't discuss them here) which led to the formulation of the two other definitions of dimension (chronologically, dim appeared in Lebesgue's papers before the appearance of ind). The latter took the place of ind in more general topological spaces.

To avoid possible misunderstandings, we wish to emphasize that, in spite of its name, the *fractal Hausdorff dimension*, now very fashionable in statistical physics, has nothing in common with the notion of *topological dimension* discussed in the present paper, for it is *not* a topological invariant.

In the sequel, whenever discussing Euclidean spaces and their subsets, we will write dim X in place of ind X.

It is clear that the definition of dimension does not single out any dimension (with the possible exception of the dimension of the empty set and the infinite dimension). This being so, it is natural to expect that the number of constructs will grow with the dimension of the space. A space of large dimension can be expected to have many relations and connections among its subsets, and the latter are likely to be marked by great complexity. In brief, we can expect that the greater the dimension, the greater the constructional possibilities. But one must also expect increased difficulties to be faced by the mathematician deprived of the assistance of intuition. On the whole, our expectations—and fears—are correct. This is confirmed by a host of theorems in geometry, in linear algebra, in the theory of manifolds, in the theory of differential equations on manifolds, and so on.

But is it *always* the case that increased dimension implies increased complexity and an increased supply of structures? The most obvious counterexample is that of vector spaces. And there are other, rather weighty counterexamples to this assertion as well as problems in which dimension plays a very subtle role. The rest of this paper is an account of these counterexamples. Their detailed analysis will enable us to answer in a mathematically reasonable manner the question in the title of this paper: *Does mathematics distinguish certain dimensions of spaces*?

3. CLASSIFICATION OF REGULAR POLYHEDRA. We begin with a problem which a mathematician is likely to dismiss as a mere curiosity, but which serves as the most elementary counterexample to the assertion that a space of large dimension is richer in structural possibilities than a space of small dimension. Specifically, we consider the classification of regular polyhedra and their higher- (and lower-) dimensional analogues in Euclidean spaces.

A complete definition of a polyhedron is too complicated to be given here. For our purposes it suffices to define a polyhedron as the intersection of a finite

number of halfspaces. A regular polyhedron is a polyhedron whose faces are congruent polygons and one with the additional property that the number of faces at each vertex is the same. Informally, we could say that a regular polyhedron is a polyhedron which looks the same from every direction. The ancients knew that there are exactly five regular polyhedra (Platonic solids), namely, the tetrahedron, the hexahedron (the cube), the octahedron, the dodecahedron, and the icosahedron (see [**6**], [**7**]).

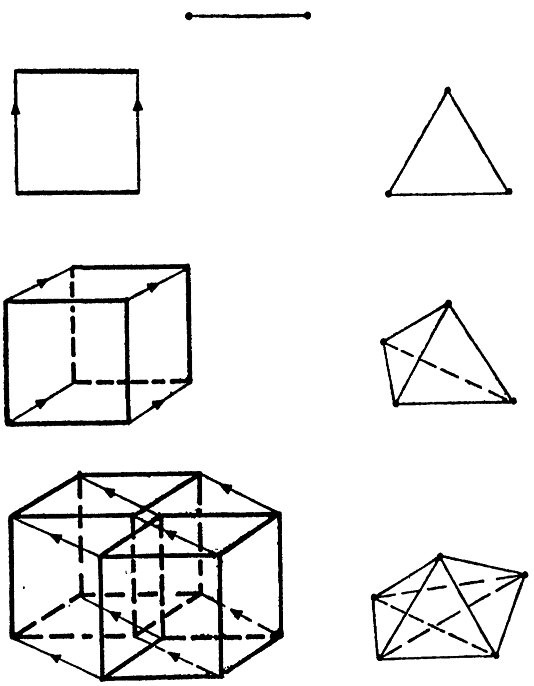


Figure 1. Lower- and higher-dimensional analogues of a cube (on the left) and a tetrahedron (on the right). From the figure we can read off the procedure for forming higher-dimensional cubes from lower-dimensional ones: a square (a 2-dimensional cube) is obtained from two appropriately located segments (1-dimensional cubes) by appropriately joining their vertices (their endpoints), a cube is obtained from two appropriately located squares by appropriately joining vertices, a hypercube is obtained from two cubes, etc. The analogues of a tetrahedron—simplexes—are the smallest convex sets containing respectively three noncollinear points (a triangle), four noncoplanar points (a tetrahedron), five points not in the same space (a 4-dimensional simplex), etc.

The role of regular polyhedra in the plane is played by regular polygons. The Greek mathematicians knew that their number is infinite. It is easy to define analogs of regular polyhedra in 4-dimensional space—the role of faces is played by ordinary 3-dimensional regular polyhedra. It turns out that there are just six such "regular cells" [**6**], [**7**]. What is surprising is that a Euclidean space of dimension 5 or higher contains just three regular cells! One of them is the n-dimensional analogue of a cube and is called a hypercube.

In order to describe the two remaining regular cells we introduce the notion of a p-dimensional face of an n-dimensional (not necessarily regular) polyhedron, $p \leq n - 1$, which is a "face of a face." For example, for $n = 3$ we have: $p = 0$

vertices, $p = 1$ edges, $p = 2$ ordinary faces of a polyhedron. It is possible to show that the number of p-dimensional faces of a hypercube is $2^{n-p}\binom{n}{p}$. We also introduce the notion of a simplex, which will be needed later as well. We take a Euclidean space of sufficiently large dimension and in it $n + 1$ points that do not lie on an $(n - 1)$-dimensional hyperplane (if $n = 3$ the four points are not coplanar). We now join the points of each pair by means of straight lines and obtain in this way an n-dimensional polyhedron whose vertices are the initial $n + 1$ points; the segments joining the vertices determine the p-dimensional faces of the polyhedron, $1 \leq p \leq n - 1$. This polyhedron is called an n-dimensional simplex. For $n = 0$ the simplex is a point, for $n = 1$ a segment, for $n = 2$ a triangle, for $n = 3$ a tetrahedron, etc. Now the second regular cell for dimension $n \geq 5$ is the n-dimensional regular simplex; it has $\binom{n+1}{p+1}$ p-dimensional faces. The third and last regular cell is the dual of the hypercube; its duality is reflected in the fact that the number of its p-dimensional faces is equal to the number of $(n - p)$-dimensional faces of the hypercube.

It would seem that the number of regular cells should increase with increasing dimension. But this is not the case. The number in question has the fixed value 3 for all $n \geq 5$. In this sense the dimensions 2, 3, and 4—and especially 2—are distinguished.

4. CLASSIFICATION OF MANIFOLDS. Our primary reason for considering the classification of manifolds is not to supply another counterexample to the assertion stated earlier. Rather, it is to illustrate the singular situations that turn up in spaces of different dimensions. The simplicity and considerable generality of the notion of a manifold has made it a leading concept of modern mathematics. One of the key advantages of manifolds is that it is possible to give their global description using methods characteristic of Euclidean spaces. Specifically, the technique created for flat spaces can be used for "curved" objects.

We will be mostly concerned with the notion of a topological manifold. An n-dimensional manifold is a Hausdorff space every point of which has a neighborhood U homeomorphic to all of $\mathbb{R}^n$ or to one of its open subsets. The mapping φ that takes the neighborhood U to an open subset of $\mathbb{R}^n$ is called a coordinate system, the pair (U, φ) is called a chart, and the value $\varphi(p) = (x^1, \dots, x^n)$ of φ at a point $p \in U$ is referred to as the coordinates of p with respect to this chart. Let (U, φ) and (U, ψ) be two charts with the same domain but with different coordinate systems. Then the composition $\varphi \circ \psi^{-1}$ maps a certain open set in $\mathbb{R}^n$ onto another open set in $\mathbb{R}^n$. The definition of a topological manifold requires this composition to be a homeomorphism in $\mathbb{R}^n$. If, in addition, we require $\varphi \circ \psi^{-1}$ and its inverse to be infinitely differentiable, then we obtain the notion of a smooth manifold [**2**].

Physicists work most of the time with smooth manifolds, familiar from courses in analysis or in tensor calculus; a prototype of such a manifold is any smooth non-self-intersecting surface in Euclidean space. Here the term "smooth surface" is somewhat misleading, for "smoothness," i.e., the existence of a tangent plane, is not necessary for a surface to be a smooth manifold (a similar remark applies to higher-dimensional manifolds). A close look at the definition shows that a surface as rich in edges as the surface of a cube is also a smooth manifold! Indeed, this surface is homeomorphic to a sphere, and every topological space homeomorphic to a smooth manifold can be given the structure of a smooth manifold. However, it

is convenient to stick to certain accepted images, and, when talking about differentiable manifolds, to have in mind objects such as an hyperboloid, a sphere, a torus, etc.

Let us go back to topological manifolds. A natural and very important problem is that of their classification. This means giving a sequence of nonhomeomorphic manifolds of the same dimension such that every manifold of this dimension is homeomorphic to one of the manifolds in the sequence. In the case of dimension 1 things are simple: every manifold is homeomorphic to a circle or to a straight line, so that the required sequence has just two elements. 2-dimensional manifolds, or surfaces, were classified at the beginning of the 20th century (in the case of dimensions 2 and higher one usually classifies only compact manifolds without a boundary, and these are the only manifolds we consider in this paper); here the sequence of representative manifolds is infinite. (More precisely, the sequence splits into two branches. The elements of one branch are connected sums of finitely many toruses, i.e., product spaces $S^1 \times S^1$, and the elements of the other branch are connected sums of finitely many projective planes.) As for 3-dimensional manifolds, the situation is as yet unclear; we do not know if these are classifiable. In 1982 William Thurston gave a partial—and very incomplete—classification of these manifolds (see **[36]** and the popular article **[8]**). We do not know what further progress is possible here.

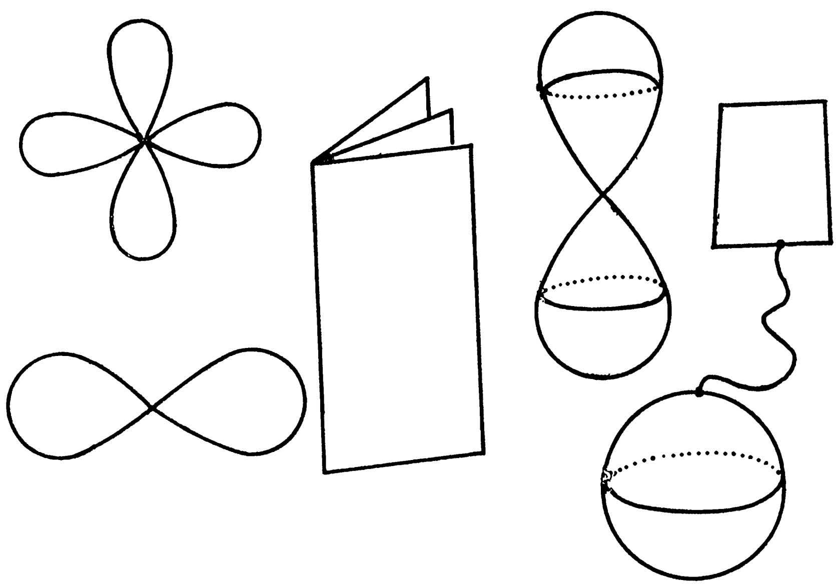


Figure 2. The sets in this figure are not manifolds. This is due to the presence of branch points or branch segments or to local changes of dimension.

Matters are clear for topological manifolds of dimension 4 and higher: in 1958 A.A. Markov proved that their classification is impossible **[9]**, **[10]**. We will try to explain why this is so. Every smooth compact manifold can be triangulated, i.e., cut into simplexes (recall that simplexes of dimension 1, 2, and 3 are segments, triangles, and tetrahedra respectively). This implies the possibility of associating with a manifold a table listing its simplexes, their faces, and the ways in which they are connected; this table is a kind of code for the manifold. It is clear that such a code is not a unique "proof of identity" of the manifold it represents, for it is

possible to cut up a given manifold in different ways and thus to associate with it different codes. This means that we would need a universal algorithm for deciding whether or not two given codes represent the same manifold (of course, the same up to a homeomorphism). If such an algorithm existed, then we could use it to divide all the codes into equivalence classes, with codes in each class corresponding to the same manifold. Markov's theorem is a nonexistence theorem: for 4-dimensional (and for higher-dimensional) manifolds there is no algorithm that would permit us to decide if two given codes correspond to homeomorphic manifolds. Thus the question of classification of manifolds is doubly hopeless. Not only is there no method of classification, but even if somebody were inspired and produced a sequence of nonhomeomorphic manifolds that exhausted all the possibilities, this result would be useless because we have no effective (i.e., universal, and requiring only a finite number of steps) way of checking which element of the sequence of representative manifolds a given manifold is homeomorphic to. In other words, even if we had a classificatory sequence of manifolds, the classification of a concrete manifold would require the carrying out of infinitely many comparison operations between this manifold and the successive elements of the sequence. We note that even if it turned out that 3-manifolds (this is a standard abbreviation) are likewise not classifiable, this impossibility would be different from that for manifolds of dimension four and up. This is so because 3-manifolds behave differently than their higher-dimensional "relatives."

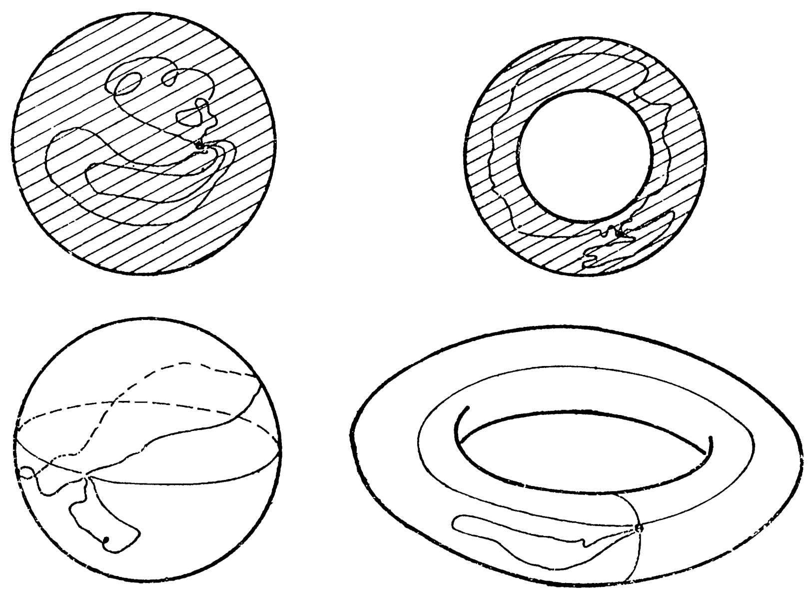


Figure 3. A disk and a sphere are simply connected, for their fundamental groups are trivial (a closed loop on either of them can be shrunk to a point). A ring and a torus are not simply connected, but their fundamental groups are different. The fundamental group of a ring is isomorphic to $\mathbb{Z}$ and the fundamental group of the torus is isomorphic to $\mathbb{Z} \oplus \mathbb{Z}$. In the case of the ring the fundamental group is generated by closed paths that loop the hole and by their integral multiples. In the case of the torus we have meridians and parallels and their multiples.

Wherein lies the difference? A detailed answer to this question would require the presentation of a number of facts from algebraic topology. This we cannot do here. What we can do is describe a certain deep difference between 3-manifolds

and 4-manifolds. The difference is that given a finitely generated group it is possible to construct a 4-manifold whose fundamental group is the given group, whereas for 3-manifolds this construction is in general impossible. More specifically, it has been shown that there are groups that cannot be the fundamental group of any 3-manifold [**9**]. One such is the direct sum $\mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}$ of four copies of $\mathbb{Z}$.

We see that topological manifolds of lowest dimension behave in a "singular" manner, in the sense that they can be classified. In this respect the dimension 4 plays the role of a limiting case.

It might seem that classification is a manifestation of excessive pedantry on the part of topologists whose love of order compels them to try to systematize the abundance of manifolds. But recently this classification has become of intense interest to physicists. The reason is that the topology of manifolds determines the possible interactions of the elementary components of matter. We give some relevant facts.

Until recently it was generally assumed that the elementary particles are pointlike. If so, then their history in spacetime is represented by worldlines. In classical accounts the interactions of particles are represented by intersecting or ramified worldlines. In quantum accounts they are represented graphically in a similar way by means of Feynman diagrams. A single line is a manifold, but an object in the form of an X or a Y is not, for a branch point, or a point of intersection, constitutes a singularity. This means that 1-dimensional topology, and in particular 1-dimensional manifolds, are useless for the description of interactions. The situation changes radically if the elementary objects are stringlike. Closed strings or loops, regarded as more interesting than open strings (segments), determine in spacetime worldtubes shaped like the surface of an asymmetric cylinder whose axis is a timelike line. We illustrate an interaction in which two strings collide and become one by means of a diagram that resembles a pair of trousers: two cylinders merging into one. The common feature of these and of more complex interactions is that the worldtubes of free and interacting strings are "proper" 2-manifolds without singular points. If a worldtube has the topology of a cylinder then the string is free. Every interaction results in a change in the topology. Topologically inequivalent worldtubes correspond to different interactions. This means that the topological classification of manifolds characterizes the possible Feynman diagrams of interacting strings (see [**34**]).

This reasoning carries over to higher-dimensional elementary objects. 2-dimensional membranes describe in spacetime 3-dimensional "worldsolids," and the difference between free and interacting membranes is reflected in the different topologies of their "solids." It is generally thought that one should consider strings and ignore membranes and higher-dimensional objects. The arguments in favor of such a position are purely technical and are based on certain formal properties of mathematical strings; their physical sense is unclear. We recall that it is impossible to classify 4-manifolds, and that the problem of classification of 3-manifolds offers difficulties that have not been surmounted thus far. This implies the impossibility of classification of Feynman diagrams corresponding to dimension 4 and our present inability to classify the Feynman diagrams corresponding to dimension 3. Is it possible that these facts offer the first serious argument in favor of strings and against membranes and other objects? The thought that there may be so intimate a connection between topology and physics is truly fascinating.

PROBLEMS AND SOLUTIONS

Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**

with the collaboration of Paul T. Bateman, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Underwood Dudley, Michael A. Filaseta, Ira M. Gessel, Bart Goddard, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Robert Israel, Murray S. Klamkin, Daniel J. Kleitman, Fred Kochman, Frederick W. Luttmann, Frank B. Miles, Richard Pfiefer, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before April 30, 1998; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.*

PROBLEMS

10620. *Proposed by James Propp, Massachusetts Institute of Technology, Cambridge, MA.* A *digraph* on a vertex set V is a subset $A \subseteq \{(v, w) : v, w \in V, v \neq w\}$ and is called *strongly connected* if it is possible to get from any vertex a to every other vertex e by a finite succession of arcs $(a, b), (b, c), \ldots, (d, e)$ in A. For $n \geq 1$, let E_n (respectively O_n) denote the number of strongly connected digraphs on the vertex set $V = \{1, 2, \ldots, n\}$ with an even (respectively odd) number of arcs. Show that $E_n - O_n = (n-1)!$ for all $n \geq 1$.

10621. *Proposed by Harold G. Diamond and Bruce Reznick, University of Illinois, Urbana-Champaign, IL.* Let $F(x)$ denote Cantor's singular function; that is, the unique non-decreasing function on $[0, 1]$ such that, if $x = \sum_{j=1}^{\infty} 2\epsilon_j/3^j$ with $\epsilon_j \in \{0, 1\}$, then $F(x) = \sum_{j=1}^{\infty} \epsilon_j/2^j$. It is clear by symmetry that $\int_0^1 F(x)\,dx = 1/2$. Prove that

$$\int_0^1 (F(x))^2\,dx = \frac{3}{10} \qquad \text{and} \qquad \int_0^1 (F(x))^3\,dx = \frac{1}{5}.$$

More generally, evaluate $\int_0^1 (F(x))^n\,dx$ for every positive integer n.

10622. *Proposed by M. N. Deshpande, Nagpur, India.* Find infinitely many triples (a, b, c) of positive integers such that a, b, c are in arithmetic progression and such that $ab + 1$, $bc + 1$, and $ca + 1$ are perfect squares.

10623. *Proposed by Roy Barbara, Lebanese University, Fanar, Lebanon.* Let $P = \{1, 2, 3, \ldots\}$ and let $|$ be the usual divisibility relation on P. For any $S \subseteq P$ and $n \in P$, $S + n = \{s + n : s \in S\}$.

(**a**) Can one construct a subset S of P such that the poset $(S, |)$ is isomorphic to $(P, |)$, $(S + 1, |)$ is isomorphic to $(P, \leq)$, and $(S + 2, |)$ is isomorphic to $(P, |)$?

(**b**) For which integers $n \geq 1$ can one find a subset T of P such that $(T, |)$, $(T + n, |)$, and $(P, |)$ are isomorphic posets?

10624. *Proposed by William F. Trench, Trinity University, San Antonio TX.* Suppose that $a_0 > a_1 > a_2 > \cdots$ and $\lim_{n\to\infty} a_n = 0$. Define

$$S_n = \sum_{j=n}^{\infty}(-1)^{j-n}a_j = a_n - a_{n+1} + a_{n+2} - \cdots.$$

Show that $\sum a_n S_n < \infty$ if and only if $\sum a_n^2 < \infty$.

10625. *Proposed by Olaf Krafft and Martin Schaefer, Technical University Aachen, Aachen, Germany.* For $x > 0$ and $n \in \mathbb{N}$, define

$$a_n = \sum_{i=0}^{2^{n-1}} \binom{2^n}{2i} x^i \bigg/ \sum_{i=0}^{2^{n-1}-1} \binom{2^n}{2i+1} x^i.$$

Evaluate $\lim_{n\to\infty} a_n$.

10626. *Proposed by Florian Luca, Syracuse University, Syracuse, NY.* For a positive integer k, the number of positive integers less than k that are relatively prime to k is denoted $\phi(k)$.
(a) Show that if m and n are relatively prime positive integers, then $\phi(5^m - 1) \neq 5^n - 1$.
(b)* Find all positive integers m, n such that $\phi(5^m - 1) = 5^n - 1$.

SOLUTIONS

Comparison Almost Everywhere

10372 [1994, 274]. *Proposed by Paul R. Chernoff and Jacob Feldman, University of California Berkeley, CA.* Let $\langle f_n\rangle_1^\infty$ be a sequence of non-negative integrable functions on the unit interval $[0, 1]$. Write $\int_0^1 f_n(x)\,dx = c_n$ and suppose that $\sum c_n < \infty$.
(a) Suppose also that $\sum \sqrt{c_n} < \infty$. Show that there is a convergent series of non-negative terms a_n such that, for almost all $x \in [0, 1]$, $f_n(x) \le a_n$ for all sufficiently large n.
(b) Show that the conclusion of (a) may fail if $\sum \sqrt{c_n} = \infty$.

Solution of **(a)** *by the proposers.* Set $a_n = \sqrt{c_n}$. The set $E_n = \{x\colon\ f_n(x) > a_n\}$ has Lebesgue measure $m(E_n) \le a_n$. Also, since $\sum a_n < \infty$, we have $\lim\limits_{N\to\infty} \sum_{n\ge N} a_n = 0$. Accordingly,

$$\lim_{N\to\infty} m\left(\bigcup_{n\ge N} E_n\right) = 0.$$

Hence $m(E) = 0$, where $E = \bigcap_{N=1}^{\infty} \bigcup_{n\ge N} E_n$. If $x \notin E$ then there is some integer $N = N(x)$ such that $x \notin \bigcup_{n\ge N} E_n$; this means that $x \notin E_n$ for all $n \ge N$, so $f_n(x) \le a_n$.

Solution of **(b)** *by Robert B. Israel, University of British Columbia, Vancouver, B. C., Canada.* Suppose $\sum_n c_n < \infty$ but $\sum_n \sqrt{c_n} = \infty$. We construct a sequence of nonnegative integrable functions f_n on $[0, 1]$ with $\int_0^1 f_n(x) = c_n$ for which there are no terms a_n that satisfy the requirements.

Let $s_n = \sum_{j=1}^{n} \sqrt{c_j}$. Let $\pi\colon \mathbb{R} \to [0, 1)$ be defined by $\pi(x) = x \bmod 1$, and let $A_n = \pi\big([s_{n-1}, s_n]\big)$. Then $m(A_n) = \sqrt{c_n}$ if $c_n \le 1$ (which is true for all but finitely many n), and every point of $[0, 1)$ is in infinitely many A_n. Take $f_n(x) = \sqrt{c_n}$ for $x \in A_n$ and $f_n(x) = 0$ otherwise. Suppose $\langle a_n\rangle$ is a nonnegative sequence with $\sum a_n < \infty$. Let $S = \{n : a_n < \sqrt{c_n}\}$ and $T = \{n : a_n \ge \sqrt{c_n}\}$. We have $\sum_{n\in T} m(A_n) \le \sum_{n\in T} a_n < \infty$. Therefore, almost every x in $[0, 1)$ is in only finitely many A_n with $n \in T$. This means that for almost every $x \in [0, 1)$ there are infinitely many $n \in S$ for which $x \in A_n$, i.e., it is not true that $f_n(x) \le a_n$ for all sufficiently large n.

Editorial comment. It should have been said more clearly that the desired property is: for almost all $x \in [0, 1]$ there is a positive integer N *depending on* x such that $f_n(x) \leq a_n$ for all $n \geq N$.

Kent Merryfield solved the more general problem: Suppose $p, q \in [1, \infty)$ and $s = pq/(p+q)$. Let $\langle f_n \rangle$ be a sequence of nonnegative functions with $\left(\int_0^1 f_n^p\right)^{1/p} = c_n$. If $\sum c_n^s < \infty$, then there is a sequence $\langle a_n \rangle$ with $\sum a_n^q < \infty$ such that for almost all x, $f_n(x) \leq a_n$ for all sufficiently large n. Furthermore, this may fail whenever $\sum c_n^s = \infty$. The original problem corresponds to $p = q = 1$ and $s = 1/2$.

Solved also by F. Galvin & P. J. Szeptycki, R. Holzsager, T. Jager, J. H. Lindsey II, O. P. Lossers (The Netherlands), L. E. Mattics, K. G. Merryfield, K. Schilling, R. B. Tucker, and J. Vinson.

Another Perspective on Napoleon's Theorem

10415 [1994, 912]. *Proposed by Edward Kitchen, Santa Monica, CA.* Let $\mathcal{A}$ be a triangle whose centroid is at the origin. Choose $k \in \mathbb{R}$, $k > 1$, and dilate one of the *Napoleon triangles* of $\mathcal{A}$ by a factor of $-k$ and the other by a factor of $k/(1-k)$. Prove that $\mathcal{A}$ is (simultaneously) perspective with both dilated triangles.

Solution by J. C. Binz, University of Bern, Bern, Switzerland. Let the triangle be given in the complex plane by the numbers $2a_1, 2a_2, 2a_3$ (counterclockwise) with $a_1 + a_2 + a_3 = 0$, and let $d = i\sqrt{3}/6$. Then the vertices of the dilated Napoleon triangles are

$$b_i = -k\big(a_{i+1} + a_{i+2} + d(a_{i+1} - a_{i+2})\big) = k\big(a_i - d(a_{i+1} - a_{i+2})\big)$$

and

$$c_i = \big(-k/(1-k)\big)\big(a_i + d(a_{i+1} - a_{i+2})\big),$$

where $a_4 = a_1$ and $a_5 = a_2$. Now, $b_i - 2a_i = (k-2)a_i - kd(a_{i+1} - a_{i+2}) = e_i$ and $c_i - 2a_i = \big(1/(1-k)\big)e_i$. Since $(b_i - 2a_i)/(c_i - 2a_i) = (1-k) \in \mathbb{R}$, the triples $2a_i, b_i, c_i$ are collinear. Let G_i denote the corresponding lines. The complex equations of the lines G_i are

$$\operatorname{Im}\big((z - 2a_i)\overline{e_i}\big) = \operatorname{Im}\big(z\overline{e_i} + 2k\overline{d}a_i(\overline{a_{i+1}} - \overline{a_{i+2}})\big) = 0,$$

where $\operatorname{Im}(z)$ stands for the imaginary part of z. The common point p of G_1 and G_2 lies on the line H with the equation $\operatorname{Im}\big((z - 2a_1)\overline{e_1}\big) + \operatorname{Im}\big((z - 2a_2)\overline{e_2}\big) = 0$. Since $\overline{e_1} + \overline{e_2} + \overline{e_3} = 0$, this equation can be written

$$\operatorname{Im}\big(-z\overline{e_3} + 2k\overline{d}(a_1\overline{a_2} - \overline{a_1}a_2 + \overline{a_3}a_2 - \overline{a_3}a_1)\big) = 0.$$

Since $\operatorname{Im}\big(\overline{d}(a_1\overline{a_2} - \overline{a_1}a_2)\big) = 0$ and $\operatorname{Im}\big(\overline{da_3}(a_2 - a_1)\big) = -\operatorname{Im}\big(\overline{d}a_3(\overline{a_1} - \overline{a_2})\big)$, we obtain finally the equation

$$\operatorname{Im}\big(z\overline{e_3} + 2k\overline{d}a_3(\overline{a_1} - \overline{a_2})\big) = 0$$

for H. Therefore H is G_3 and the three lines G_1, G_2, and G_3 are concurrent.

Editorial comment. The solver also noted that the proof could be generalized to allow the Napoleon triangles to be replaced by triangles formed by the vertices of any set of similar isosceles triangles built on the sides of the triangle. This construction leads to the Kiepert hyperbola (see R. H. Eddy and R. Fritsch, The conics of Ludwig Kiepert: A comprehensive lesson in the geometry of the triangle, *Math. Mag.* 67 (1994) 188–205). If the hyperbola is parameterized by the base angle ϕ of the isosceles triangle, then dilation by a factor λ leads to the point on the hyperbola whose parameter θ satisfies $\tan\theta/\tan\phi = 3\lambda/(2+\lambda)$.

Solved also by J. Anglesio (France), M. Benedicty, J. Ferrer (Spain), J. Fukuta (Japan), C. G. Petalas & T. P. Vidalis (Greece), R. Tauraso (Italy), NSA Problems Group, and the proposer.

A Paucity of Diophantine Solutions

10445 [1995, 359]. *Proposed by Alan J. Gross, Medical University of South Carolina, Charleston, SC, and Hong Zhang, Indiana-Purdue University, Fort Wayne, IN.* Note that $5^2 + 5 + 2 = 2^5$. Are there any other positive integers a and b with $a^b + a + b = b^a$?

Solution by the late J. Sutherland Frame, Michigan State University, East Lansing, MI. No. Setting $b = 1$ forces $a = 0$, which is not positive. Setting $b = 2$ yields only $a = 5$, since $2^a > (a+1)^2$ for $a \geq 6$. Hence we may assume that $b \geq 3$. It is also easy to show that there are no solutions with $a = 1$ or $a = 2$. The inequality $a^b < b^a$ implies that $a^{1/a} < b^{1/b}$. Since $x^{1/x}$ decreases with increasing x when $x > e$, we obtain $a > b \geq 3$.

Let $f(x) = x^b + x + b - b^x$. We have $f(b) = 2b > 0$. We claim that $f(b+1) < 0$ and that f is decreasing for $x > b \geq 3$, so its only root above 3 lies between b and $b+1$ and cannot be an integer. To prove that $f(b+1) < 0$, observe that for $b \geq 3$,

$$b^{b+1} = eb^b + (b-e)b^b > (1 + 1/b)^b b^b + .28b^b > (b+1)^b + (2b+1).$$

For $x > b \geq 3$, we have observed that $x^b < b^x$. Thus

$$f'(x) = bx^{b-1} + 1 - (\ln b)b^x = b^x\left(\frac{b}{x}\frac{x^b}{b^x} + b^{-x} - \ln b\right) < b^x\left(1 + \frac{1}{27} - \ln 3\right) < 0.$$

Solved also by J. Anglesio (France), R. Barbara (Lebanon), D. Beckwith, M. R. Burke & L. Sweet (Canada), R. J. Chapman (U. K.), J. Christopher, R. B. Eggleton, K. Foltz & N. Rosenberg, Z. Franco, S. M. Gagola Jr., R. A. Groeneveld, G. A. Heuer, J. Kholdi, N. Komanda, W. C. Lang, J. H. Lindsey II, J. H. van Lint (The Netherlands), L. E. Mattics, C. A. Minh, E. D. Onstott, A. Pedersen (Denmark), R. E. Prather, R. Stong, A. A. Tarabay (Lebanon), A. N. 't Woord (The Netherlands), W. C. Wu (China), C. Y. Yildirim (Turkey), Anchorage Math Solutions Group, NSA Problems Group, Oklahoma State Problems Group, and the proposers.

Conjugate Generators

10449 [1995, 360]. *Proposed by Frank Schmidt, Arlington, VA.* For which n can the symmetric group S_n be generated by two *conjugate* permutations?

Solution by Fred Galvin, University of Kansas, Lawrence, KS. This holds for all n. We prove the stronger statement that if a group G is generated by elements a and b such that $a^2 = b^{2m+1} = 1$ for some integer m, then G is generated by the two conjugate elements $f = ab$ and $g = afa$. The proof is that $a = f(gf)^m$ and $b = (gf)^{m+1}$.

For $n > 1$ it is well known that S_n is generated by the transposition $a = (12)$ and the cycle $c = (12 \ldots n)$. It is also generated by a and the $(n-1)$-cycle ac. Thus in applying the result in the previous paragraph we take $b = c$ if n is odd and $b = ac$ if n is even.

Editorial comment. The work of most solvers led to a generator that was a cycle of maximal even length. For $n \geq 5$, John H. Lindsey II gave an example that was a product of a transposition and a maximal disjoint cycle of odd length. Stephen M. Gagola Jr. asked whether examples could be found if the cycle structure was subject to additional restrictions. Two necessary conditions are (1) there are an odd number of even cycles, and (2) there are fewer than $n/2$ fixed points. These conditions have been important in general work on generators of the symmetric group (see section 53 of Daniel Gorenstein (editor), Reviews on Finite Groups, AMS, 1974). In addition, Gagola mentioned (3) the element is not of order 2 if $n > 3$, which is necessary because two elements of order 2 generate a dihedral group.

Solved also by M. Brodie, R. J. Chapman (U. K.), S. M. Gagola Jr., J. H. Lindsey II, G. R. Robinson (England), R. Stong, D. B. Tyler, P. Venzke, A. N. 't Woord (The Netherlands), Anchorage Math Solutions Group, NSA Problems Group, Oklahoma State Problems Group, WMC Problems group, and the proposer.

A Surprising Family of Continued Fractions

10457 [1995, 464]. *Proposed by Henry Cohn, Massachusetts Institute of Technology, Cambridge, MA.* Determine the simple continued fraction of $(F_{10n+1}/F_{10n})^5$, where F_k denotes the kth Fibonacci number.

Solution by B. M. M. de Weger, Krimpen aan den IJssel, The Netherlands. The answer is

$$\frac{F_{10n+1}^5}{F_{10n}^5} = [11,\ldots,11,10,1,1,\frac{F_{20n+1}-6}{5},1,17,11,\ldots,11,10,1,4,11,\ldots,11],$$

where the strings of 11's have lengths $2n-1$, $2n-2$, and $2n-1$, respectively. Note that indeed $(F_{20n+1}-6)/5 \in \mathbb{Z}$. We prove this by translating continued fraction expansions into matrix products and using the following lemma.

Lemma. If $(p/q) > 1$ is a rational number in lowest terms, and $a_0,\ldots,a_n$ are positive integers with $a_n \geq 2$, then $(p/q) = [a_0,\ldots,a_n]$ if and only if integers c, d exist such that

$$\begin{pmatrix} p & c \\ q & d \end{pmatrix} = \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix}\cdots\begin{pmatrix} a_n & 1 \\ 1 & 0 \end{pmatrix}.$$

Proof. For $0 \leq m \leq n$, let (p_m/q_m) in lowest terms be defined by $(p_m/q_m) = [a_m,\ldots,a_n]$. Also let

$$\begin{pmatrix} r_m & c_m \\ s_m & d_m \end{pmatrix} = \begin{pmatrix} a_m & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a_{m+1} & 1 \\ 1 & 0 \end{pmatrix}\cdots\begin{pmatrix} a_n & 1 \\ 1 & 0 \end{pmatrix}.$$

Then $p_n = r_n = a_n$ and $q_n = s_n = 1$. By

$$[a_m,\ldots,a_n] = a_m + \frac{1}{[a_{m+1},\ldots,a_n]} \quad \text{and} \quad \begin{pmatrix} r_m & c_m \\ s_m & d_m \end{pmatrix} = \begin{pmatrix} a_m & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} r_{m+1} & c_{m+1} \\ s_{m+1} & d_{m+1} \end{pmatrix},$$

we have the same recurrence relations for (p_m, q_m) as for (r_m, s_m). □

In view of the lemma, it suffices to show for some integers c, d that

$$\begin{pmatrix} F_{10n+1}^5 & c \\ F_{10n}^5 & d \end{pmatrix} = \begin{pmatrix} 11 & 1 \\ 1 & 0 \end{pmatrix}^{2n-1}\begin{pmatrix} 10 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}^2\begin{pmatrix} \frac{F_{20n+1}-6}{5} & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$
$$\times\begin{pmatrix} 17 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 11 & 1 \\ 1 & 0 \end{pmatrix}^{2n-2}\begin{pmatrix} 10 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 4 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 11 & 1 \\ 1 & 0 \end{pmatrix}^{2n-1}$$

In expanding the matrix product, we use the identities $F_{20n+1} = F_{10n}^2 + F_{10n+1}^2$, $F_{10n-5} = -8F_{10n} + 5F_{10n+1}$, $F_{10n-10} = 89F_{10n} - 55F_{10n+1}$, $F_{10n-15} = -987F_{10n} + 610F_{10n+1}$, and

$$\begin{pmatrix} 11 & 1 \\ 1 & 0 \end{pmatrix}^k = \frac{1}{5}\begin{pmatrix} F_{5k+5} & F_{5k} \\ F_{5k} & F_{5k-5} \end{pmatrix}.$$

The needed matrix identity now reduces to two identities involving F_{10n} and F_{10n+1}:

$$5F_{10n+1}^5 = F_{10n}^5 + 2F_{10n}^4F_{10n+1} + 2F_{10n}^3F_{10n+1}^2 + 4F_{10n}^2F_{10n+1}^3 + F_{10n}F_{10n+1}^4 + 2F_{10n+1}^5$$
$$+F_{10n}^3 + F_{10n}^2F_{10n+1} + 2F_{10n}F_{10n+1}^2 + 3F_{10n+1}^3,$$

$$5F_{10n}^5 = 2F_{10n}^5 - F_{10n}^4F_{10n+1} + 4F_{10n}^3F_{10n+1}^2 - 2F_{10n}^2F_{10n+1}^3 + 2F_{10n}F_{10n+1}^4 - F_{10n+1}^5$$
$$-3F_{10n}^3 + 2F_{10n}^2F_{10n+1} - F_{10n}F_{10n+1}^2 + F_{10n+1}^3.$$

These can be proved by inserting $F_k = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^k - \left(\frac{1-\sqrt{5}}{2}\right)^k\right)$ and expanding the products.

Solved also by J. Anglesio (France), R. J. Chapman (U. K.), and the proposer.

Solitaire on the Circle

10459 [1995, 553]. *Proposed by David Beckwith, Sag Harbor, NY.* A game is played with n disks ($n \geq 3$), each having a black face and a red face. Initially, the n disks are arranged in a circle showing a random pattern of black and red faces. A move consists of taking away a black disk (i.e., one with its black face exposed) and inverting its neighbors (if any). The resulting gap is not closed up, so the remaining disks do not acquire new neighbors. The goal is to remove all the disks. For which initial patterns is this possible?

Solution I by Glenn G. Chappell, Southeast Missouri State University, Cape Girardeau, MO, and Chris Hartman, University of Alaska, Fairbanks, AK. It is possible if and only if the inital pattern has a nonzero even number of black disks. Note that the first move (including the flips) changes the parity of the number of black disks. Hence it suffices to prove that contiguous disks with a gap at each end can all be removed if and only if the line has an odd number of black disks or is empty.

The statement is trivial for the empty string; we proceed by induction on n. Suppose first that the number of black disks is odd. We may choose a black disk D having an even number (possibly zero) of black disks to each side. Removing D creates two (possibly empty) shorter lines of disks. If either side is nonempty, completing the move by flipping the neighbor of D creates a line on that side with an odd number of black disks. By the induction hypothesis, the remaining disks on both sides can be removed.

Now suppose that the nonempty line has an even number of black disks. If it has no black disks, then nothing can be removed. Otherwise, each black disk D has an odd number of black disks to one of its sides. Removing D and flipping its neighbor on that side creates a smaller line with an even number of black disks which, by the induction hypothesis, cannot be removed. Thus there is no first disk to remove that permits success.

Solution II by O. P. Lossers, University of Technology, Eindhoven, The Netherlands. To prove the same result, we label disks with the integers modulo n. When there is a successful removal procedure, we define a bijection $\pi\colon \mathbb{Z}_n \to \mathbb{Z}_n$ by letting $\pi(i) = j$ when disk i is removed at stage j. A disk initially red has been inverted once when it is removed; a disk initially black has been inverted zero or two times. Thus an original black disk is removed before both neighbors or after both neighbors, while an original red disk is removed between its neighbors. Thus the periodic extension of i ($\pi(n) = \pi(0)$, etc.) has a local extremum at i precisely when i is initially black. These local extrema are alternately maxima and minima, so there are an even number of them.

For sufficiency, we construct such a bijection when the number of black disks is nonzero and even. We begin by giving the black disks distinct real labels that are alternately positive and negative as we traverse the circle. Between a pair of black disks, we assign labels to the red disks so that all labels are distinct and the labels are monotone on this interval. Finally, we assign j to the disk that has received the jth smallest number.

Editorial comment. Randall Maddox of Pepperdine University used Multimedia Toolbook to create a version of the game that runs in Microsoft Windows 3.1. He suggested that playing the game in this environment would be fun while leading to the discovery of the solution. He has offered to make this application available to anyone who wants it.

Solved also by S. F. Barger, M. Benedicty, W. C. Calhoun, D. Callan, M. Černe (Slovenia), R. J. Chapman (U. K.), Y. S. Chen, S. M. Gagola Jr., W. Gasarch & P. Godfrey, R. Holzsager, R. D. Hurwitz & S. Hurwitz, L. Keener & S. Walters (Canada), P. G. Kirmser, C. Lindsey, J. H. Lindsey II, J. H. van Lint (The Netherlands), H. Loimer (Austria), R. B. Maddox, R. Martin (Germany), M. D. Meyerson, B. Mixon, J. H. Nieto (Venezuela), A. Onshuus (Colombia), A. Pedersen (Denmark), K. Rebman, G. Rice, M. Rierson, A. J. Schwenk, K. Sonnichsen, R. Stong, J. R. Stoughton, S. Szabo (Hungary), A. A. Tarabay & B. B. Ghalayini (Lebanon), Y. Wang, M. Woltermann, A. N. 't Woord (The Netherlands), Anchorage Math Solutions Group, Con Amore Problem Group (Denmark), NSA Problems Group, Prague Problem Solution Group (Czech Republic), Circulo de Matématicas – Universidad de Los Andes (Colombia), and the proposer.

A Demanding Currency

10465 [1995, 554]. *Proposed by Paul K. Stockmeyer, College of William and Mary, Williamsburg, VA.* As the Minister of Finance of a newly independent country, you must design a new currency: a sequence $d_1 < d_2 < d_3 < \cdots$ of positive integers with $d_1 = 1$, to be the denominations of various coins and bills. Although you are authorized to create an infinite number of denominations, the legislature has passed some laws restricting your choices.
Rule 1: There must be a bound b on the number of items needed for any payment.
Rule 2: The "denomination density", $\lim_{k\to\infty} k/d_k$, must be zero.
Rule 3: Repeatedly choosing the largest denomination less than or equal to the amount remaining to be paid (the greedy algorithm) always leads to the use of the minimal number of items to pay any amount.

Can you design a currency meeting these rules?

Solution by Ellen Hertz, National Highway Traffic Safety Administration, Washington, DC. There is no such currency. Rule 2 implies that $\{d_{k+1} - d_k\}$ is unbounded. By Rule 1, we can select an amount n that requires the maximum number, b, of items. Select an index k such that $d_{k+1} - d_k > n$. Consider the amount $d_k + n$. By Rule 3, we can start with d_k to obtain the minimum number of items needed to pay $d_k + n$. However, doing so leaves the amount n remaining, which requires b additional items.

Editorial comment. Several readers observed that any two of the rules can easily be satisfied. The set of all integers satisfies rules 1 and 3, and a geometric progression satisfies rules 2 and 3. An example satisfying rule 1 with $b = 4$ and rule 2 is given by $d_k = k^2$. One can satisfy rule 1 with $b = 2$ and rule 2 by having d_k list in increasing order the set of numbers whose ternary expansion omits the digit 2. Furthermore, while rule 3 with the usual interpretation fails for this example, if one applies the greedy algorithm separately to each ternary digit of the amount to be paid, one obtains a number in the set for which the balance is also in the set.

Solved also by S. F. Barger, W. C. Calhoun, G. G. Chappell & C. M. Hartman, F. M. Djorup, P. Freyd, S. M. Gagola Jr., T. Hwa, L. Keener (Canada), J. H. Lindsey II, O. P. Lossers (The Netherlands), R. Martin (Germany), C. H. Montenegro (Colombia), G. Myerson (Australia), A. J. Schwenk, D. C. Terr, Y. Wang, D. R. Witte, G. J. Woeginger (The Netherlands), NSA Problems Group, Prague Problems Group (Czech Republic), and the proposer.

The Number of Positive Semidefinite 0,1-Matrices

10481 [1995, 840]. *Proposed by Frank Schmidt, Arlington, VA.* Let $f(n)$ denote the number of positive semidefinite n by n matrices whose entries are 0 or 1. Let $g(n)$ denote the number of positive definite n by n matrices whose entries are 0 or 1. Evaluate $f(n)$ and $g(n)$.

Solution by David Callan, University of Wisconsin, Madison, WI. The answers are $g(n) = 1$ and $f(n) = B(n+1)$, where $B(n)$ is the number of partitions of an n-element set (the nth *Bell number*). The definitions permit only symmetric matrices. The determinant criterion for a real symmetric matrix to be positive definite [positive semidefinite] is that its principal submatrices all have positive [nonnegative] determinant. Thus a positive definite 0,1-matrix has 1's on the diagonal (consider the 1 by 1 submatrices) and 0's off the diagonal (consider the 2 by 2 submatrices). Hence $g(n) = 1$.

Every square matrix of 1's is positive semidefinite; its 1 by 1 principal submatrices have determinant 1 and the others have determinant 0. Thus A is positive semidefinite if there is a permutation matrix P such that $P^{-1}AP$ is a direct sum of a zero matrix and matrices of all 1's. Such a matrix is specified by partitioning $\{0, 1, \ldots, n\}$; the nonzero elements in the block containing 0 are the indices of the rows and columns in the 0 matrix, and each other block gives the rows and columns in a block of 1's. In this way we obtain $B(n+1)$ such matrices.

We claim that these are the only positive semidefinite 0,1-matrices. Given a positive semidefinite 0,1-matrix A, let $J = \{i : a_{ii} = 1\}$. For $i \notin J$, the 2 by 2 submatrices involving $a_{ii} = 0$ show that all entries in the ith row and ith column are 0. Define a relation $\sim$ on J by setting $i \sim j$ if and only if $a_{ij} = 1$. Reflexivity and symmetry are immediate for $\sim$. If $\sim$ is not transitive, then there exist $i, j, k \in J$ such that $a_{ij} = a_{jk} = 1$ but $a_{ik} = 0$. This yields a 3 by 3 principal submatrix B that is all 1's except for one off-diagonal pair of 0's. Since the determinant of such a matrix is -1, this cannot happen. We conclude that $\sim$ is an equivalence relation, which expresses A in the form described above.

Editorial comment. The characterization of positive semidefinite 0,1-matrices appears in R. A. Horn, The theory of infinitely divisible matrices and kernels, *Trans. Amer. Math. Soc.* 136 (1969) 269-286 and also as exercises 4 and 5 on page 457 of R. A. Horn & C. Johnson, Topics in Matrix Analysis, Cambridge University Press, 1991.

Solved also by S. Byrd & R. L. Smith, O. Krafft (Germany), S.-O. Troschke (Germany), and the proposer.

The Exponent Isn't Variable

10508 [1996, 266]. *Proposed by Jesús Ferrer, Universidad de València, Burjasot, Spain.* Let $f, g : \mathbb{R} \to \mathbb{R}$ be two infinitely differentiable functions with g analytic. Show that if, for each point $x \in \mathbb{R}$, there is a positive integer $\mu(x)$ such that $f(x) = g(x)^{\mu(x)}$, then f is a constant power of g, i.e., there is a fixed positive integer n such that $f(x) = g(x)^n$ for all $x \in \mathbb{R}$.

Solution by Nicholas Passell and Alexander Smith, University of Wisconsin-Eau Claire, Eau Claire, WI. If g is constant the problem is trivial. Otherwise by analyticity of g the set $T = \{x : g(x) = 0, 1, \text{ or } -1\}$ consists of isolated points with no accumulation point. Except at the points of T, the function $\mu(x) = (\ln|f(x)|) / (\ln|g(x)|)$ is continuous. Hence $\mu(x)$ is constant between any two of the isolated points of T. Suppose $t \in T$ and that $f(x) = g(x)^{n_1}$ on $[a_1, t]$ and $f(x) = g(x)^{n_2}$ on $[t, a_2]$. Since $f(x)$ is infinitely differentiable, its left and right derivatives of all orders at t are equal. Thus $g(x)^{n_1}$ and $g(x)^{n_2}$ are analytic and their derivatives of all orders at t agree. Hence $g(x)^{n_1} = g(x)^{n_2}$ for all x. Looking at any point $x \notin T$ we see $n_1 = n_2$. Thus the exponent may be chosen constant throughout $\mathbb{R}$.

Solved also by R. J. Chapman (U. K.), J. Cobb, J. Hejcman (Czech Republic), S. S. Kim (Korea), C. A. Kumar & P. V. S. P. Saradhi (India), J. H. Lindsey II, L. E. Mattics, M. McKinzie, A. Pedersen (Denmark), R. Stong, T. V. Trif (Romania), GCHQ Problems Group (U. K.), NSA Problems Group, and the proposer.

Random Distribution on the Sphere

10516 [1996, 347]. *Proposed by Donald A. Darling.* Let (X, Y, Z) be three random variables such that $\alpha X + \beta Y + \gamma Z$ is uniformly distributed in the interval $[-1, 1]$ for every set of three *direction cosines*, i.e., numbers with $\alpha^2 + \beta^2 + \gamma^2 = 1$. Show that $X^2 + Y^2 + Z^2 = 1$ with probability one.

Solution I by Richard A. Groeneveld and Stephen B. Vardeman, Iowa State University, Ames, IA. Let $V = X^2 + Y^2 + Z^2$. We show that $E(V) = 1$ and $\mathrm{Var}(V) = 0$, so $V = 1$ with probability 1. Using $\alpha = 1$, $\beta = 1$, and $\gamma = 1$ successively, each of X, Y, and Z has the distribution of U, where the latter is a uniform random variable on $[-1, 1]$. Thus $E(X^2) = E(Y^2) = E(Z^2) = E(U^2) = 1/3$, and so $E(V) = E(X^2 + Y^2 + Z^2) = 1$.

Also $(X + Y)/\sqrt{2}$ and $(X - Y)/\sqrt{2}$ both have the distribution of U, so

$$\frac{E\left((X+Y)^4\right) - E\left((X-Y)^4\right)}{4} = 2E(X^3Y + XY^3) = 0.$$

One has $E(X^4) = E(Y^4) = 1/5$, and

$$\frac{E\left((X+Y)^4\right)}{4} = \frac{E(X^4 + 4X^3Y + 6X^2Y^2 + 4XY^3 + Y^4)}{4} = \frac{1}{5}.$$

Using $E(X^3Y + XY^3) = 0$ yields $E(X^2Y^2) = 1/15$, and similarly, $E(X^2Z^2) = E(Y^2Z^2)$ $= 1/15$. Hence,

$$\begin{aligned}
\operatorname{Var}(V) &= E\left((X^2 + Y^2 + Z^2)^2\right) - 1 \\
&= E(X^4 + Y^4 + Z^4 + 2X^2Y^2 + 2X^2Z^2 + 2Y^2Z^2) - 1 \\
&= 3\left(\frac{1}{5}\right) + 6\left(\frac{1}{15}\right) - 1 = 0,
\end{aligned}$$

which completes the proof.

Examination of the proof shows that in order that $X^2 + Y^2 + Z^2 = 1$ with probability 1, it is sufficient that $\alpha X + \beta Y + \gamma Z$ have the distribution of U for the 9 sets of direction cosines $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, $\left(1/\sqrt{2}, \pm 1/\sqrt{2}, 0\right)$, $\left(1/\sqrt{2}, 0, \pm 1/\sqrt{2}\right)$, and $\left(0, 1/\sqrt{2}, \pm 1/\sqrt{2}\right)$.

Solution II by Robin J. Chapman, University of Exeter, Exeter, U. K. We use the fact that the probability distribution of a triple of random variables (X, Y, Z) is determined by its *characteristic function* $\chi(u, v, w) = E(\exp(i(uX + vY + wZ)))$. But if $(u, v, w) = r(\alpha, \beta, \gamma)$ with $\alpha^2 + \beta^2 + \gamma^2 = 1$, then

$$\begin{aligned}
\chi(u, v, w) &= E(\exp(ir(\alpha X + \beta Y + \gamma Z)) = \frac{1}{2}\int_{-1}^{1} e^{irt}\,dt \\
&= \frac{\sin r}{r} = \frac{\sin\sqrt{u^2 + v^2 + w^2}}{\sqrt{u^2 + v^2 + w^2}}.
\end{aligned}$$

Now let (X_1, Y_1, Z_1) be a triple of random variables uniformly distributed according to area on the unit sphere. If χ_1 is its characteristic function, then χ_1 is spherically symmetric. Calculation using spherical coordinates gives

$$\chi_1(0, 0, w) = \frac{1}{4\pi}\int_0^{\pi}\int_0^{2\pi} e^{iw\cos\theta}\sin\theta\,d\varphi\,d\theta = \frac{\sin w}{w}.$$

By spherical symmetry, $\chi_1 = \chi$ and so (X, Y, Z) and (X_1, Y_1, Z_1) have the same distribution. Hence $X^2 + Y^2 + Z^2 = 1$ with probability one.

Similar arguments work for other numbers of variables. If (X, Y) are random variables with $\alpha X + \beta Y$ uniform on $[-1, 1]$ for all α and β with $\alpha^2 + \beta^2 = 1$, then the distribution of (X, Y) is uniquely determined, and the joint probability density function of (X, Y) is $(1/2\pi)(1 - x^2 - y^2)^{-1/2}$. But if $n \geq 4$ and $(X_1, \ldots, X_n)$ are random variables with $\sum_{j=1}^{n} \alpha_j X_j$ uniform on $[-1, 1]$ whenever $\sum_{j=1}^{n} \alpha_j^2 = 1$, then $X_1^2 + X_3^2 + X_4^2 = X_2^2 + X_3^2 + X_4^2 = 1$ with probability one. Hence $X_1^2 = X_2^2$ with probability one, but this cannot occur since (X_1, X_2, X_3) is uniformly distributed over the unit sphere. It follows that this result cannot be extended to more than three variables.

Editorial comment. Using the methods of Solution I, Mark Pinsky showed that for the analogous n-dimensional problem, $\operatorname{Var}(X_1^2 + \cdots + X_n^2) = (n-3)(n+5)/15$; so for $n > 3$ the norm *cannot* be almost surely constant. Solution II establishes the stronger conclusion that (X, Y, Z) itself is uniformly distributed over the surface of the sphere. (The truth of the converse was noted by several solvers.) Can one find (X, Y, Z) that satisfies the hypothesis

for the nine directions specified in Solution I but is not uniform on the sphere? The reference A. Renyi, On projections of probability distributions, *Acta Math. Sci. Hungar.* **3** (1952) 131–142 was noted by Groeneveld and Vardeman; there is also the paper by our proposer: D. A. Darling, On a problem of Renyi, *Period. Math. Hungar.* **3** (1973) 5–7. Renyi proved a more general result: A distribution in $\mathbb{R}^3$ is determined by its projections on every straight line through the origin.

Solved also by P. J. Fitzsimmons, E. Hertz, J. H. Lindsey II, L. E. Mattics, K. G. Merryfield, W. A. Newcomb, S. Northshield, N. Passell, M. A. Pinsky, K. Schilling, R. Stong, GCHQ Problems Group (U. K.), and the proposer.

An Excluded Circle

10517 [1996, 347]. *Proposed by Jean Anglesio, Garches, France.* Let $\triangle ABC$ be a triangle and let H be its orthocenter and I its incenter. If W is the point such that $\overrightarrow{HW} = 4\,\overrightarrow{HI}$ and $R = 2\sqrt{2}\,|HI|$, show that none of the vertices A, B, or C is in the interior of the circle with center W and radius R.

Solution by Joel Rosenberg, University of Michigan, Ann Arbor, MI. As usual, let us write a, b, and c for the lengths of sides BC, CA, and AB, respectively. Let O be the circumcenter of $\triangle ABC$, and let ρ be its circumradius. We analyze the problem by using vectors centered at O and writing P for the vector OP. We also write (α, β, γ) for $\alpha A + \beta B + \gamma C$. Observe that $(B - C) \cdot (B + C) = \rho^2 - \rho^2 = 0$, and cyclically, so we conclude that $H = A + B + C = (1, 1, 1)$. Also, for any P,

$$P = \left(\frac{\text{Area}(\triangle PBC)}{\text{Area}(\triangle ABC)}, \frac{\text{Area}(\triangle PCA)}{\text{Area}(\triangle ABC)}, \frac{\text{Area}(\triangle PAB)}{\text{Area}(\triangle ABC)}\right),$$

so in particular $I = \big(a/(a+b+c), b/(a+b+c), c/(a+b+c)\big)$ and

$$W = 4I - 3H = \left(\frac{a-3b-3c}{a+b+c}, \frac{-3a+b-3c}{a+b+c}, \frac{-3a-3b+c}{a+b+c}\right).$$

The definition of circumcenter gives $A \cdot A = B \cdot B = C \cdot C = \rho^2$, and the law of cosines gives $A \cdot B = \rho^2 - c^2/2$, $B \cdot C = \rho^2 - a^2/2$, and $C \cdot A = \rho^2 - b^2/2$. Thus if $P = (u, v, w)$ and $Q = (x, y, z)$, then

$$P \cdot Q = \begin{pmatrix} u & v & w \end{pmatrix} \begin{pmatrix} \rho^2 & \rho^2 - (c^2/2) & \rho^2 - (b^2/2) \\ \rho^2 - (c^2/2) & \rho^2 & \rho^2 - (a^2/2) \\ \rho^2 - (b^2/2) & \rho^2 - (a^2/2) & \rho^2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

We conclude that

$$\begin{aligned} \left|\overrightarrow{HI}\right|^2 &= (I - H) \cdot (I - H) = \frac{4\rho^2(a+b+c) - a^3 - b^3 - c^3 - abc}{a+b+c}, \\ \left|\overrightarrow{WA}\right|^2 &= (A - W) \cdot (A - W) \\ &= \frac{36\rho^2(a+b+c) - 9a^3 - 12b^3 - 12c^3 + 4b^2c + 4bc^2 - 16abc}{a+b+c}. \end{aligned}$$

Now we write K for Area$(\triangle ABC)$, and recall the law of sines $K = abc/4\rho$ and Heron's formula $16K^2 = (a+b+c)(a+b-c)(c+a-b)(b+c-a)$. Thus we can write

$$\left|\overrightarrow{WA}\right|^2 = \frac{f(a,b,c)}{16K^2} \qquad \text{and} \qquad \left|\overrightarrow{HI}\right|^2 = \frac{g(a,b,c)}{16K^2},$$

where f and g are sixth degree homogeneous polynomials in a, b, and c. Using *Mathematica*, we find that

$$|\overrightarrow{WA}|^2 - 8|\overrightarrow{HI}|^2 = \frac{f(a,b,c) - 8g(a,b,c)}{16K^2}$$
$$= \frac{\left(a^3 - 2a^2b - 2a^2c - ab^2 + 4abc - ac^2 + 2b^2 - 2b^2c - 2bc^2 + 2c^3\right)^2}{16K^2} \geq 0.$$

Thus $|\overrightarrow{WA}| \geq 2\sqrt{2}\,|\overrightarrow{HI}|$, and similarly $|\overrightarrow{WB}| \geq 2\sqrt{2}\,|\overrightarrow{HI}|$, $|\overrightarrow{WC}| \geq 2\sqrt{2}\,|\overrightarrow{HI}|$, so none of the points A, B, C lies within the circle centered at W with radius $2\sqrt{2}\,|\overrightarrow{HI}|$.

Editorial comment. All four solutions were quite computational; two used Mathematica.

Solved also by M. Benedicty, V. Schindler (Germany), and the proposer.

An Integral Giving Euler's Constant

10524 [1996, 427]. *Proposed by Jean Anglesio, Garches, France.* Show that

$$\lim_{u\to\infty} \int_{1/u}^{u} \left(\frac{\sin x}{x} - \cos x\right) \frac{\ln x}{x}\, dx = 1 - \gamma,$$

where γ is Euler's constant.

Solution by Donald A. Darling, Newport Beach, CA. Transform

$$I(u) = \int_{1/u}^{u} \left(\frac{\sin x}{x} - \cos x\right) \frac{\ln x}{x}\, dx\,.$$

by integrating by parts:

$$I(u) = \left(1 - \frac{\sin x}{x}\right) \ln x \Big|_{1/u}^{u} - \int_{1/u}^{u} \left(1 - \frac{\sin x}{x}\right) \frac{dx}{x}$$
$$= \left(1 - \frac{\sin u}{u}\right) \ln u - \left(1 - u \sin\frac{1}{u}\right) \ln\frac{1}{u} - \int_{1/u}^{u} \left(1 - \frac{\sin x}{x}\right) \frac{dx}{x}\,.$$

Use $\int_{1/u}^{u} (1+x)^{-1}\, dx = \ln u$ to write the previous equation as

$$I(u) = -\sin u \frac{\ln u}{u} + \left(1 - u \sin\frac{1}{u}\right) \ln u + \int_{1/u}^{u} \left(\frac{\sin x}{x} - 1 + \frac{x}{1+x}\right) \frac{dx}{x}\,.$$

Since $1 - u\sin(1/u) \sim 1/(6u^2)$ as $u \to \infty$, the first two terms on the right go to zero, and the integral over $[0, \infty)$ is absolutely convergent, so that

$$I(u) \to \int_0^\infty \left(\frac{\sin x}{x} - \frac{1}{1+x}\right) \frac{dx}{x} = 1 - \gamma$$

by a standard representation of Euler's constant—see, for example, I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, formula 8.367(8) or 3.781(1).

Editorial comment. Gaston Gonnet noted that the *Maple* computer algebra system immediately returns the asserted value for the limit.

Solved also by Z. Ahmed (India), D. Borwein (Canada), D. Bradley (Canada), R. J. Chapman (U. K.), E. Deutsch, M. Golomb, G. H. Gonnet (Switzerland), R. A. Groeneveld, K.-W. Lau (Hong Kong), M. Omarjee (France), R. Richberg (Germany), P. G. Rooney, N. S. Thornber, T. V. Trif (Romania), M. Vowe (Switzerland), C. Y. Yildirim (Turkey), GCHQ Problems Group (U. K.), NSA Problems Group, and the proposer.

REVIEWS

Edited by **Underwood Dudley**
Mathematics Department, De Pauw University, Greencastle, IN 46135

Knots and Surfaces: A Guide to Discovering Mathematics. By David W. Farmer and Theodore B. Stanford. American Mathematical Society, 1996, vii + 101, $19.00.
Knots and Surfaces. By N. D. Gilbert and T. Porter. Oxford University Press, 1994, xi + 268, $75.00 hardbound, $34.95 softcover.

Reviewed by William D. Dunbar

The two books under review, which I'll refer to as [**FS**] and [**GP**], have the same title, but are very different in content and style. [**FS**] is, in its own words, "a guide to discovering mathematics... suitable for a one semester course at the beginning undergraduate level [with] no prerequisites." The discussion is driven by a succession of "Tasks" (over 150 in the three main chapters on networks, surfaces and knots), which include both calculations and short proofs of the "explain why" sort. The six-color theorem for maps on a sphere is proven in the first chapter, orientable surfaces (with or without boundary) are classified in the second chapter, and the invariance of linking number and 3-colorability under Reidemeister moves is proven in the third chapter (asserting but not proving that this implies invariance under ambient isotopy). The final chapter consists of a list of ideas for projects or term papers. Some are generalizations of material presented earlier, some address the use of knots and surfaces in art, and some are analyses of games played on surfaces (e.g., tic-tac-toe on a torus). I feel that [**FS**] would serve well as the basis of an independent study course, in which the student would work through the tasks in a journal, subject to periodic review by the instructor (to monitor the proof-writing). The writing is clear and engaging, and the tasks should be effective at setting a reasonable pace. One of the most memorable tasks for me was to label a map of the continental United States, which has been rotated and distorted homeomorphically so that the states are polygons that fit together into a rectangle (territorial waters are included in the state, so Michigan has one component). Being from New England, I knew to start with Maine.

If [**FS**] might better be titled "Graphs and Surfaces and Knots," then [**GP**] is really "Groups and Knots and Surfaces and Graphs." In its own words, it is "based on a third-year undergraduate course" at a British university. The first two chapters discuss knot diagrams and several of the knot polynomials defined for knot diagrams (Jones, Kauffman, HOMFLY, Alexander). The next two chapters introduce the basic notions of topology and apply them to the classification of surfaces (orientable or not, with or without boundary). Chapter 5 proves that the genus of a connected sum of two knots is the sum of the genera of the summands, and outlines a proof of the prime factorization of knots. Also the HOMFLY polynomial of a connected sum is proven to be the product of the polynomials of the summands.

Group theory begins in Chapter 6, with Cayley graphs, construction of free groups, group presentations, Tietze transformations, cyclic decomposition of finitely generated abelian groups, pushouts of groups, and the Wirtinger & Dehn presentations associated to a knot diagram. Chapter 7 switches to graphs, especially as embedded in surfaces, and closes by giving without proof the minimum genus surface in which the complete graph K_n on n vertices can be embedded. Fox derivatives are used in Chapter 8 to give another facet of the Alexander polynomial. The following three chapters introduce the fundamental group of a topological space, prove the Seifert-Van Kampen theorem and apply it to surfaces and to knots. The final chapter is a general discussion of covering spaces, which are then applied to graphs, ending with the Schreier formula for the rank of a subgroup of given index in a free group of given rank.

I am not enthusiastic about **[GP]** as a textbook, for two basic reasons: its narrative line is weak, and it is not well edited. By the former, I mean that the enormous investment in combinatorial group theory in the middle of the book does not seem to pay off adequately in terms of a better understanding of knots or surfaces. In particular, I found the ending of the text quite abrupt. As for the latter, I'll give a few examples of things that either mystified or annoyed me:

(1) On page 52, a knot is defined as a map $S^1 \to \mathbf{R}^3$, but on page 53, a theorem asserts that two knots are ambient isotopic if and only if they have isotopic (unoriented) diagrams, which appears to be false for non-invertible knots such as 8_{17}. Since this definition gives an implicit orientation to all knots, the later use of the term "oriented knots" on page 96 is confusing.

(2) On page 94, the fact that the degree of the Alexander polynomial of a knot is no more than twice the genus of the knot is called "strange," but it follows immediately from $\Delta(x) = \det(V - xV^t)$, where V is a Seifert matrix for the knot (with $2g$ rows and columns, where g is the genus of the Seifert surface).

(3) On page 112 and throughout, group homomorphisms are misspelled as "homeomorphisms."

(4) On page 180, the reader is advised that "now would be a good time to look up the definitions" of an integral domain and a unique factorization domain. This statement is immediately followed by a string of definitions, including those two.

(5) Repeated explanations of Lebesgue's lemma and changes in the notation for paths make it appear that Chapters 9 and 10 were written by different people without consideration of the transition. $\overline{w}$ is used both for the reverse of the path w (on page 231) and for something quite different (on page 228).

By contrast, I've taught a topology course from C. Livingston's book **[L]**, and feel that it gives a clearer picture of the wide variety of knot invariants, how two invariants can be related by an inequality or independent of each other, and how having several methods to calculate the same thing (such as the Alexander polynomial) can be intriguing as well as useful. The algebra is quickly applied; the discussion of permutations supports both the development of the concept of the knot group (via labelling edges in a knot diagram "consistently" with permutations) and also the later results on linking numbers of periodic knots with their axes of rotational symmetry. **[L]** ends with an open question, that of finding a knot that cannot be distinguished from the unknot by, say, the HOMFLY polynomial. Indeed, there are many references to open problems throughout the book, which reinforce the point that mathematics is an evolving subject, a point that undergraduates need not only to hear, but to see.

REFERENCE

[L] Livingston, Charles, *Knot Theory*, MAA, 1993.

Division of Natural Sciences and Mathematics
Simon's Rock College of Bard
Great Barrington, MA 01230
wdunbar@simons-rock.edu

The Book of Numbers. By John Horton Conway and Richard K. Guy. Springer-Verlag, 1996, 320, $29.95.

Reviewed by **Andrew Bremner**

Am I the only person ever to have claimed a pineapple as deduction against income tax? The arrival of Conway and Guy's *Book of Numbers* may well mean that others follow suit when classroom teachers discover the pedagogical virtues of using the fruit to demonstrate the occurrence in nature of the Fibonacci numbers. What a delight it is to turn the pages of this book, being simultaneously entertained and enlightened by these masters of arithmetic mathesis. The planning of such a book displays enormous conceit, and bringing it to realisation is a remarkable achievement. Readers must enjoy discovering for themselves quite how successful the authors have been. With "numbers" as the theme of the book, the volume could well have been arbitrarily large; yet in practice it is quite a slim tome of just ten chapters (short of fourteen, one notes, the infinity of Borges).

It is in fact admirably restrained in size; it is also beautifully produced by Springer, and comfortably priced to boot. What more could one ask? There is a clear readership amongst the devoted followers of Martin Gardner, and *Scientific American*'s Mathematical Games columns. Readers familiar with the two volume *Winning Ways* by Berlekamp, Conway, and Guy [**1**], will undoubtedly recognize inimitable matters of style common to both works, including many whimsically delightful illustrations. But the material of the *Book of Numbers* is intentionally far more accessible to the lay mathematician than that of *Winning Ways*, and can be perused with pleasure and profit by pappous schoolboy and pompous professor. There is truly something here for everyone.

The first chapter starts with the language of number, from the hypothesis that "Hickory, Dickory, Dock" is a corruption of some rustic counting sequence for "Eight, Nine, Ten", to the syntax of number names in different languages. As illustration, a table of cardinals in Welsh is given, where we learn for instance that "eighteen" occurs as "two times nine" and "fifty" as "half a hundred". (Those of us who have slogged through classical Hawaiian will recall the intrigue at "seventy" translating as "forty with thirty remainder" despite the regular formations of the other multiples of ten.) There follow individual ruminations based on the numbers between one and a hundred. So we find the familiar, the French "quinzaine" denoting the same period of time as the English "fortnight"; and the less familiar —"punch", a drink with five ingredients, from the Hindi word for five.

This whole section begs a parlour game with the aim of finding some literary reference to the integers between one and a hundred (maybe the Bible should be disbarred for this purpose. John Buchan provides an easy start; I suspect a search of *Tristram Shandy* would also prove fruitful.) The visual equivalent exists on celluloid in Greenaway's "Drowning by Numbers": the viewer becomes aware that the numbers from one to one hundred are occurring in sequence in the film, with an abrupt ending at the appearance of 100.

We progress to historical number systems (and discover how Caesar dealt with fractions: for instance, the five spots on a die is the Roman quincunx, the symbol for five uncia, or five twelfths. Of course, there are still today twelve ounces in both the troy and the apothecary's pound). Later in the book is a consequent discussion of tablet number 7289 from the Yale Babylonian Collection, an astonishing base 60 computation of the square root of 2. Its accuracy is such that the value is correct to six decimal places. Also shown is a photograph of the tablet Plimpton 322 which in Babylonian cuneiform appears to be a table of Pythagorean triangles (here duly completed by the authors, in Babylonian of course).

The second chapter includes Patterns Providing Pretty Proofs, with some geometrically inspired formulations of series summations. The Ackermann numbers are described, being the sequence $1, 4, 7625597484987, \ldots$ whose fourth term is so staggeringly immense that the cerebellum quivers merely at the thought of trying to comprehend it. Yet this is only a springboard for the authors' "chained-arrow" numbers; these in turn are of such magnitude as guaranteed to leave your brainbox smoking.

The following chapter concentrates on sequences, and in particular methods for determining the rule of formation of a sequence of integers. Little is assumed, so that binomial coefficients and Pascal's triangle are introduced from first principles. But there is an interesting discussion of Difference Fans and of Number Walls, of which I choose the latter to illustrate here with an example. If standard differencing techniques do not reveal the rule behind a sequence, then the sequence cannot be of polynomial type. To detect sequences of exponential type (linear recurrence relations) one forms a Number Wall: below a row of ones place the terms of the sequence, and then compute further rows in the wall using the rule that for each cross of five bricks,

$$(\text{centre-brick})^2 = (\text{north-brick})(\text{south-brick}) + (\text{east-brick})(\text{west-brick})$$

That is, for the cross

$$\begin{array}{ccc} & N & \\ W & C & E \\ & S & \end{array}$$

$C^2 = NS + EW$. So for the Fibonacci sequence, for example, we develop the wall

$$\begin{array}{ccccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \\ 0 & 1 & 1 & 2 & 3 & 5 & 8 & 13 & 21 & 34 & \cdots \\ & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & \cdots \\ & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \end{array}$$

If the sequence is genuinely exponential, then ultimately a row of zeroes will appear, and the number of rows between the ones and zeroes gives the length of the recurrence relation, which is then easy to compute. What is so intriguing about this construction is the surely non-obvious fact that every brick will be computed as

an *integer*. To prove all the properties of the Number Wall will clearly afford excellent exercise. In more substantial vein, I apply the Number Wall technique to the Shallit sequence (see [2]): a_{n+1} is the least integer such that

$$\frac{a_{n+1}}{a_n} > \frac{a_n}{a_{n-1}},$$

with $a_0 = 8$, $a_1 = 55$. The sequence is thus $8, 55, 379, 2612, 18002, \ldots$. We construct the wall

1	1	1	1	1	1	1	1	1	$\cdots$
8	55	379	2612	18002	124071	855106	5893451	40618081	$\cdots$
	-7	-19	-214	-1448	-5171	-87785	-82185	-4520276	$\cdots$
		-3	7	55	-809	8515	-66185	501931	$\cdots$
			-1	-6	-36	-216	-1296	-7776	$\cdots$
				0	0	0	0	0	$\cdots$

which indicates a four term recurrence relation. Indeed, we appear to find

$$a_n = 6a_{n-1} + 7a_{n-2} - 5a_{n-3} - 6a_{n-4}.$$

But beware, O constructor murorum! It is anti-intuitive that the sequence $\{a_n\}$ be exponential, and in fact with sufficient computer-aided effort it has been verified that the linear recurrence is valid only in the range $4 \le n \le 11055$. The term a_{11056} differs by 1 from that computed from the recurrence! It is instructive to estimate the size of a_{11056}, and contemplate the fact that the bottom row of the above wall starts out with 11051 zeroes before becoming non-zero.

This is heady stuff, but we are propelled forward into Famous Families of Numbers, where the Great Arithmetician of Ulm, Johann Faulhaber, is finally given full credit for his discovery and use in the *Academiae Algebrae* (1631) of what later became known as the Bernoulli numbers. It was almost eighty years later that Bernoulli himself embarked on his extensive study of these numbers (with, it should be remarked, full due to Faulhaber). So we find Bell and Catalan, Ramanujan and Stirling, and of course Fibonacci. There is an extended discussion of the occurrence in nature of the Fibonacci numbers, from the edible sorosis to leaf phyllotaxis. A mature pineapple will display eight spirals of bracts in one direction, and thirteen in the other, clockwise and anticlockwise (a supermarket sortie found no counterexample to this rule; the lady who observed my protracted handling of every single pineapple advanced with the advice: "You should *smell* them, you know"). Similarly a sunflower head will usually display 34 and 55 spirals of seeds. The authors provide what seems a very plausible explanation for this phenomenon ("Say, bud, where do you think you are going?"). D'Arcy Thompson's admirable *On Growth and Form* was a favourite tome of the childhood bookshelf, but is sadly remiss in failing even to mention the Fibonacci sequence.

In the Primacy of Primes, we meet *inter alia* Conway's famous Prime Producing Machine, and we are given the current status on the factorization of the Mersenne and Fermat primes. The latter, $F_n = 2^{2^n} + 1$, is now known to be composite for $5 \le n \le 23$; factorizations are given for $5 \le n \le 11$, and partial factorizations for $n = 12, 13$. This is certainly very much up-to-date, with F_{10} and F_{11} having only just fallen (however the authors do miss two very recently announced Mersenne primes).

Fruitfulness of Fractions is a rag-bag collection of Farey fractions, decimal expansions of prime reciprocals, shuffles, and Pythagorean triangles. Continued fractions, which themselves might merit an entire volume, are restricted to a

mention in terms of the astronomical Metonic cycle, used to determine the Jewish calendar and the date of Easter. (What a shame Conway was not prevailed upon to add a page or two at this stage, even if not *strictly* relevant. Master of much, he is supreme in the quirks of the calendar, and I have savoured for many years his perspicacious observation that any Swede born on February 29 in the year 1696 would not have celebrated a first birthday for another 48 years!)

"Algebraic Numbers" leads to the ruler-and-compass construction of the regular polygons, and beautifully elegant constructions with appropriate diagrams are given for the 3-, 5-, 7-, 9-, 13-, and 17-gons ("But..." I hear you murmur; yes, indeed, the luxury of an angle-trisector is used for the 7-, 9-, and 13-gons.) Some problems are given whose solution will involve a specific algebraic number, for instance, that of finding a hexagon of largest area given that no two vertices are more than one unit of distance apart. Ron Graham solves the problem with a hexagon of area A, with A satisfying an irreducible polynomial of degree 10 ($A = 0.674981\ldots$; the area of a regular hexagon of unit side is $3\sqrt{3}/8 = 0.649519\ldots$). There is even a somewhat contrived problem whose solution involves the root of a polynomial of degree 71. It should perhaps be stressed that throughout the book, all these deeper results are simply quoted, and each chapter has a sizeable bibliography referring the reader to research papers where necessary, and to appropriate literature in general.

The remaining chapters include sections on imaginary and transcendental numbers, with mention of the connection between Euler's prime-producing polynomial $n^2 - n + 41$ and the fact that $e^{\pi\sqrt{163}}$ is an integer. ("But..." I hear you mutter again: you must simply go away and compute, but make sure to give yourself at least 31 significant figures.) There is also a nice discussion of Gregory numbers, where the name of Lewis Carroll arises.

Finally, in the unique style of the authors, there is a chapter on the infinite and the infinitesimal. We are shown how to add and multiply, just as if infinities occurred every day in our cheque-books. Surrealism comes into play, as it were, but it's only a game. Sorry. You just have to get hold of a copy of this book; trying to summarize adequately the 320 pages of Conway & Guy is as demanding an exercise as extracting the plums from a particularly rich Christmas pudding.

So the book has multifarious virtues: what are its faults? Being ever greedy, one can lament what the authors have chosen to omit, as much as rejoice in what they include. Here again is another parlour game, to alphabetize missing topics that perhaps merit mention: amicable numbers and Alcuin's sequence, Beatty sequences, congruent numbers... There is a slightly irritating and curious inconsistency with the chosen type-faces. Springer have chosen a rather spidery font for some of the tables, which can render the content impactless. For instance, Table 6.1 lists those repeating decimals that occur in fractions with denominator a given prime. One of the features that should leap to the eye is the equal length of the entries for a given prime: 027, 054, 081,... for the prime 37. As it stands, however, entries such as that at 53 look anything but equal in length. Yet other tables, such as Table 1.4, decimal expansions of "some of our favourite numbers", are set perfectly. Which is not in the index; at least, "perfect" is not in the index. A little unfortunate, since this was the very first item that I looked up on receiving the book (perfect numbers are indeed mentioned on pages 136–137). The index also mispaginates at least one item. Tut. However, if such be the sum total of faults, then there is not much cause for curmudgeonly grumble.

It is clear that this eclectic review can only begin to convey the pleasure that this book has provided. Throughout, the authors communicate their enthusiasm and

exuberance with great éclat. There is joy at being in the care of mathematicians who delight in the sheer *friendliness* of numbers. The lucid explanations and insights can be startling and impressive. Gounod set to music in his curiously admonitory duet "L'Arithmétique":

> ...
> Cultiver cet art salutaire
> C'est apprendre à garder son bien,
> Car, mes amis, sur cette terre,
> Sachez, qu'on a souvent affaire
> A des gens qui comptent trop bien.

How fortunate that these authors who count so much better than most of us have imparted their wisdom to the printed medium. In Japan at New Year, the takarabune is the treasure-laden ship of the Gods of Good Fortune. Here is our ship, with Conway and Guy the Bringers of Happiness.

REFERENCES

1. E. R. Berlekamp, J. H. Conway, and R. K. Guy, *Winning Ways: for your mathematical plays*, Academic Press, 1982 (2 vols.).
2. D. Boyd, Linear recurrence relations for some generalized Pisot sequences, *Advances in Number Theory* (Kingston Ontario, 1991), 333–340, Oxford Sci. Publ., Oxford University Press, New York, 1993.

Department of Mathematics
Arizona State University
Tempe, Arizona 85287
bremner@asu.edu

There is an old Armenian saying, "He who lacks a sense of the past is condemned to live in the narrow darkness of his own generation." Mathematics without history is mathematics stripped of its greatness: for, like the other arts—and mathematics is one of the supreme arts of civilization—it derives its grandeur from the fact of being a human creation

G. F. Simmons, *Differential Equations with Applications and Historical Notes*, second edition, McGraw-Hill, Inc., 1991, p. xix

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1–4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly.*

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

Mathematics Appreciation, T(13: 1). *Mathematics: Its Power and Utility, Fifth Edition.* Karl J. Smith. Brooks/Cole, 1997, xv + 588 pp, \$61.95. [ISBN 0-534-34462-3] New for this edition: more historical information; more problems asking students to explain concepts in their own words; glossary of important terms; appendices on Hindu–Arabic numbers and enumeration systems. (*Fourth Edition*, TR, January 1995.) LB

Mathematics Appreciation, S, L. *Power Play.* Edward J. Barbeau. Spectrum Ser. MAA, 1997, xi + 185 pp, \$29 (P). [ISBN 0-88385-523-2] Lots of neat stuff about powers of integers, Pythagorean triples, Pell's Equation, Catalan Conjecture, and more. Includes a collection of "interesting sets" (e.g., {7, 18} is "interesting" because $7^3 = 1+18+18^2$). Good exercises. JO

Finite Mathematics, T(13–14: 1). *Applied Finite Mathematics, Fifth Edition.* S.T. Tan. Brooks/Cole, 1997, xvii + 674 pp, \$73.95. [ISBN 0-534-95562-2] Business oriented; emphasis on linear models (systems, linear programming), finance, sets and counting, elementary probability, Markov chains and games. Many applied examples. (*Third Edition*, TR, August–September 1990.) RM

Education, P, L. *Many Visions, Many Aims, Volume 1.* William H. Schmidt, *et al.* Kluwer Academic, 1997, ix + 276 pp, \$120. [ISBN 0-7923-4436-7] A cross-national investigation of curricular intentions in school mathematics based on an analysis of textbooks and curriculum guides in almost fifty countries participating in the Third International Mathematics and Science Study (TIMSS). Key finding: variation ("noise") was pervasive "in myriad ways both small and large," dominating commonalities ("signal") across both countries and topics. LAS

Education, P, L*. *A Splintered Vision: An Investigation of U.S. Science and Mathematics Education.* William H. Schmidt, Curtis C. McKnight, Senta A. Raizen. Kluwer Academic, 1997, 163 pp, \$87. [ISBN 0-7923-4440-5] Summary report of U.S. curricula, textbooks, and instructional practice drawn from the Third International Science and Mathematics Study (TIMSS). Chief conclusion: U.S. texts include, and U.S. teachers teach, far more topics than is typical in other nations. This lack of focus, the authors assert, is due to the absence of any "coherent vision" of U.S. school mathematics. LAS

Education, S(15–18), P. *Assessment Standards for School Mathematics.* NCTM, 1995, ix + 102 pp, \$15 (P). [ISBN 0-87353-419-0] Third book in NCTM *Standards* trilogy. Assumes all students can meet high standards and that assessment is interwoven with instruction. Addresses four purposes of assessment: monitoring students' progress, making instructional decisions, evaluating students' achievement, and evaluating programs. Presents six standards: mathematics, learning, equity, openness, inferences, and coherence. Numerous examples. MW

Education, S(17–18), P. *Mathematics for Tomorrow's Young Children: International Perspectives on Curriculum.* Eds: Helen Mans-

field, Neil A. Pateman, Nadine Bednarz. Math. Educ. Lib., V. 16. Kluwer Academic, 1996, xi + 327 pp, $128. [ISBN 0-7923-3998-3] Papers from a Working Group at ICME–7 focus on concept development and conceptual change, including the influences of social and cultural environments, the roles of language and symbolism, and the interactions between teachers' curriculum decisions and their beliefs about concept formation. Strong focus on the role of social interaction in students' construction of mathematical ideas, and on similarities and differences among various constructivistic types of learning theories. MW

Education, S(17–18), P. *Approaches to Algebra: Perspectives for Research and Teaching.* Eds: Nadine Bednarz, Carolyn Kieran, Lesley Lee. Math. Educ. Lib., V. 18. Kluwer Academic, 1996, xv + 345 pp, $145. [ISBN 0-7923-4145-7] Collection of research papers on the development of algebraic thinking among secondary school students. Focus is on didactic questions such as: what students should know, what learning is prerequisite, what obstacles to learning arise, what situations promote algebraic thinking. Questions are discussed in the context of four conceptions of algebra: (1) as generalizations of numerical and geometric patterns and laws governing numerical relations; (2) as a means of solving of specific problems or classes of problems; (3) as a tool for modeling physical phenomena; and (4) as a study of the concepts of variable and function. MW

Education, S(17–18), P. *Gender and Mathematics Education.* Eds: Barbro Grevholm, Gila Hanna. Lund Univ Pr, 1995, 428 pp. [ISBN 91-7966-276-5] Proceedings of a 1993 International Commission on Mathematical Instruction (ICMI) conference. Explores the causes, manifestations, and possible solutions for gender inequities in mathematics participation and persistence. Papers address specific issues and present "snapshots" of gender issues in nearly a dozen countries. MW

History, P, L. *A Logical Journey: From Gödel to Philosophy.* Hao Wang. MIT Pr, 1996, xiv + 391 pp, $40. [ISBN 0-262-23189-1] A continuation of the author's *Reflections on Kurt Gödel* (TR, January 1988). Reports the author's conversations with Gödel on interpretation of his published work. Fascinating chapters on mind vs. machine and on Gödel's Platonism and objectivism. An example: Gödel believed both "... mathematics describes a non-sensual reality which exists independently ... of the human mind" and that his 1951 Gibbs Lecture proved this. SK

Number Theory, T(18: 1), P*. *Duality in Analytic Number Theory.* P.D.T.A. Elliott. Tracts in Math., V. 122. Cambridge Univ Pr, 1997, xviii + 341 pp, $59.95. [ISBN 0-521-56088-8] A mathematical autobiography: one of the masters explains the tools of analytic number theory, especially sieve theory, within the context of how he has used them and how he thinks about them. DB

Number Theory, T(18: 1), P. *The Hardy–Littlewood Method, Second Edition.* R.C. Vaughan. Tracts in Math., V. 125. Cambridge Univ Pr, 1997, xiii + 232 pp, $49.95. [ISBN 0-521-57347-5] Expanded to include recent results on Waring's problem, especially Wooley's result that $G(k)$ is bounded by $k \log k + k \log\log k + O(k)$. (*First Edition*, TR, March 1982.) DB

Number Theory, P. *Surfing on the Ocean of Numbers—A Few Smarandache Notions and Similar Topics.* Henry Ibstedt. Erhus Univ Pr, 1997, 75 pp, $9.95 (P). [ISBN 1-879585-57-X]

Group Theory, T(18: 1), S, P. *Automorphic Forms and Representations.* Daniel Bump. Stud. in Adv. Math., V. 55. Cambridge Univ Pr, 1997, xiv + 574 pp, $79.95. [ISBN 0-521-55098-X] Written at a level intermediate between an advanced text and a monograph. The four chapters, which can be read independently, are Modular Forms, Automorphic Forms and Representations of $GL(2, R)$, Automorphic Representations, and Representations of $GL(2)$ Over a p-adic Field. JS

Group Theory, P. *Linear Algebra and Group Theory for Physicists.* K.N. Srinivasa Rao. Wiley, 1996, 623 pp, $34.95. [ISBN 0-470-22061-9] Starts with basic group theory and linear algebra, then covers representation theory including representations of finite groups, linear associative algebras, the symmetric group, the rotation group, crystallographic point groups, and the Lorentz group. LC

Group Theory, T(18), P. *Low Rank Representations and Graphs for Sporadic Groups.* Cheryl E. Praeger, Leonard H. Soicher. Australian Math. Soc. Lect. Ser., V. 8. Cambridge Univ Pr, 1997, xi + 141 pp, $39.95 (P). [ISBN 0-521-56737-8]

Algebra, P, L*. *Finite Fields.* Rudolf Lidl, Harald Niederreiter. Ency. of Math. & Its Applic., V. 20. Cambridge Univ Pr, 1997, xiv + 755 pp, $95. [ISBN 0-521-39231-4] Reprint of an excellent resource. (1983 Addison-Wesley edition, TR, April 1987.) CEC

Algebra, T(18: 2). *Groups and Characters.* Larry C. Grove. Pure & Appl. Math. Wiley,

1997, viii + 212 pp, $54.95. [ISBN 0-471-16340-6] Focuses primarily on finite groups: character theory, transfer and splitting, Frobenius groups. Includes some computational aspects: Schreier–Sims algorithm, Todd–Coxeter coset enumeration, etc. LC

Algebra, T(18), P. *Representation Theory and Complex Geometry.* Neil Chriss, Victor Ginzburg. Birkhäuser Boston, 1997, x + 495 pp, $64.50. [ISBN 0-8176-3792-3] Overview of recent advances in representation theory—particularly for Weyl groups, $SL_n(\mathbb{C})$, and affine Hecke algebras—emphasizing the geometric aspects: Springer theory, K-theory, flag varieties, perverse sheaves. Self-contained, with many new results and new approaches. Accessible to students of Lie theory. TH

Differential Equations, P. *Frequency Methods in Oscillation Theory.* G.A. Leonov, I.M. Burkin, A.I. Shepeljavyi. Math. & Its Applic., V. 357. Kluwer Academic, 1996, xii + 403 pp, $196. [ISBN 0-7923-3896-0] Criteria for the existence of periodic solutions to ODEs. Translation of Russian original. SK

Partial Differential Equations, P. *Time Dependent Problems and Difference Methods.* Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger. Pure & Appl. Math. Wiley, 1995, xi + 642 pp, $79.95. [ISBN 0-471-50734-2] Introduction to PDEs and numerical methods for them. Directed at physical scientists and engineers. SK

Differential Geometry, P. *Geometry of Nonpositively Curved Manifolds.* Patrick B. Eberlein. Lect. in Math. Ser. Univ of Chicago Pr, 1996, vii + 449 pp, $45 (P); $90. [ISBN 0-226-18198-7; 0-226-18197-9] Presentation of recent results on manifolds of non-positive sectional curvature, leading up to a proof of the Mostow Rigidity Theorem. JO

Differential Geometry, T(18: 1). *Differential Geometry: Cartan's Generalization of Klein's Erlangen Program.* R.W. Sharpe. Grad. Texts in Math., V. 166. Springer-Verlag, 1997, xix + 421 pp, $49.95. [ISBN 0-387-94732-9] Development of differential geometry leading to the global version of Cartan connections. The exposition is nicely influenced by the history of the subject, and by the author's question of modern differential geometry: "What [has] become of the curves and surfaces?" JO

Mathematical Modeling, S(13–15), L. *Interdisciplinary Lively Application Projects (ILAPs).* Ed: David C. Arney. MAA, 1997, xii + 222 pp, $27.50 (P). [ISBN 0-88385-706-5] Eight interdisciplinary projects, akin to UMAP modules, followed by six brief articles on the philosophy of their development and use. Projects range from the economics of building a deck to energy consumption on a mountain bike, from modeling smog in Los Angeles to heavy metal contaminants in an aquifer. Prerequisites range from precalculus to advanced calculus and differential equations. Each project includes instructions for students, comments for instructors, and sample solutions. A product of the NSF-supported INTERMATH project centered at West Point. LAS

Probability, T(17), P. *A Weak Convergence Approach to the Theory of Large Deviations.* Paul Dupuis, Richard S. Ellis. Ser. in Prob. & Stat. Wiley, 1997, xvii + 479 pp, $69.95. [ISBN 0-471-07672-4] Shows how to apply weak convergence theory in a consistent way to prove large deviation results. Develops the weak convergence approach from scratch, illustrates it via examples, applies it to sophisticated models. KB

Mathematical Statistics, T(14–16). *Statistical Methods: A Geometric Primer.* David J. Saville, Graham R. Wood. Springer-Verlag, 1996, xi + 268 pp, $39.95 (P). [ISBN 0-387-94705-1] Unique geometric perspective on fundamental statistical methods. Focuses on providing insights into mathematical foundations of the methods, yet all topics introduced and motivated through scientific problems and data analyses. Graphics enhance understanding and exercises are interesting/appropriate. MK

Statistical Methods, T(18), P, L. *Sequential Estimation.* Malay Ghosh, Nitis Mukhopadhyay, Pranab K. Sen. Ser. in Prob. & Stat. Wiley, 1997, xiv + 480 pp, $69.95. [ISBN 0-471-81271-4] Classical and modern techniques in sequential estimation including parametric and nonparametric methods. Topics: shrinkage, empirical and hierarchical Bayes procedures, time-sequential estimation, empirical and hierarchical populations sampling, reliability estimations, and capture-recapture methodology leading to sequential schemes. KB

Statistical Methods, S(17–18). *Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS Examples.* Sylvie Huet, *et al.* Ser. in Stat. Springer-Verlag, 1996, ix + 154 pp, $42.95. [ISBN 0-387-94727-2] Well-written text. Includes numerous examples, theoretical constructs, and fully-worked applications. Topics include nonlinear regression and parameter estimation; interval estimation and testing; variance estimation; diagnosing model misspecifications; calibration and prediction. No exercises. MK

Statistical Methods, P. *Applied Wavelet Analysis with S-PLUS.* Andrew Bruce, Hong-Ye Gao. Springer-Verlag, 1996, xxi + 338 pp, $49.95 (P). [ISBN 0-387-94714-0] Beautiful book introduces wavelet theory through scientific applications and data analyses, including image and speech analyses. Approach is graphical/visual. Written for scientists and researchers interested in applying these techniques. No exercises. MK

Statistical Methods, T(17–18: 1), S. *The EM Algorithm and Extensions.* Geoffrey J. McLachlan, Thriyambakam Krishnan. Ser. in Prob. & Stat. Wiley, 1997, xvii + 274 pp, $59.95. [ISBN 0-471-12358-7] Excellent resource for theory and application of EM algorithm and its many variations. Applications range from simple, one-parameter multinomials to hidden Markov models, epidemiological models, and neural networks. No exercises. MK

Statistical Methods, T(16–18: 1, 2). *Modern Regression Methods.* Thomas P. Ryan. Ser. in Prob. & Stat. Wiley, 1997, xix + 515 pp, $64.95, with disk. [ISBN 0-471-52912-5] Thorough survey covers simple linear regression through nonparametric, robust, and nonlinear regression. Plenty of good exercises and data sets. Could be used for introductory regression course or for more advanced methods course. MK

Statistics, T(13–14: 1). *Statistics from Scratch: An Introduction for Health Care Professionals.* David Bowers. Wiley, 1996, vi + 180 pp, $19.95 (P). [ISBN 0-471-96325-9] Survey of introductory statistical topics for health science students. Tone makes material accessible. Includes standard introductory topics as well as stratified sampling plans, cohort studies, case-control studies, and clinical trials. Concepts stressed without mathematical detail. Relies on accessible software. Numerous exercises. MK

Theory of Computation, T(16–17: 1, 2), L. *Computability and Complexity: From a Programming Perspective.* Neil D. Jones. Found. of Computing. MIT Pr, 1997, xvi + 466 pp, $45. [ISBN 0-262-10064-9] Interesting, integrated treatment of models of computation (programming languages, TM's, lambda calculus, inference systems, etc.), and complexity, based on and motivated by notions and techniques from programming language perspective, via basic model of LISP-like programs. Nicely exploits theoretical and practical connections (self evaluators, partial evaluation). RM

Computer Science, T(16–17: 1), P. *Specification of Abstract Data Types.* Jacques Loeckx, Hans-Dieter Ehrich, Markus Wolf. Wiley, 1996, xi + 260 pp, $60. [ISBN 0-471-95067-X] Introduction to the theory of algebraic program specification. Unlike some texts in the field, this one is intended to be accessible to students unfamiliar with category theory. Includes design of a lexical analyzer as a case study. JO

Applications (Economics), P. *Prices, Cycles, and Growth.* Hukukane Nikaido. Stud. in Dynam. Econ. Sci. MIT Pr, 1996, xix + 285 pp, $40. [ISBN 0-262-14059-4] Fourteen of the author's research papers analyzing the stability of equilibria and cycles in various economic models. SK

Applications (Engineering), T(15–17: 1, 2), S, P. *Fuzzy Engineering.* Bart Kosko. Prentice Hall, 1997, xxvi + 549 pp, with disk. [ISBN 0-13-124991-6] Fuzzy systems are approximations $F : \mathbb{R}^n \to \mathbb{R}^p$ modeling an associative processor given by m rules of the form "if X is a fuzzy set A, then Y is a fuzzy set B." Approach departs from the common "linguistic" view (fuzziness modeling, how we reason with rules of thumb, etc.), to perspective of function approximation with additive fuzzy systems. Treats chaos and control, signal processing, communication, etc. RM

Applications (Physical Science), S(17–18), P. *The Sentinel Method and Its Application to Environmental Pollution Problems.* Jean-Pierre Kernévez. Math. Modelling Ser. CRC Pr, 1997, 204 pp, $69.95. [ISBN 0-8493-9630-1] Describes a method for solving parameter estimation problems for environmental pollution models for which data are incomplete. LB

Applications (Physical Science), S(11–14), L. *Mission Mathematics: Grades 9–12.* Peggy House. NCTM, 1997, vi + 121 pp, $19.95 (P). [ISBN 0-87353-436-0] An innovative collection of authentic case studies of mathematical ideas behind space science: launch windows, orbital debris, global positioning systems, elliptical orbits, communication links. Each chapter introduces its subject with extensive references to real events and data, then is followed by student assignments on reproducible worksheets. LAS

Reviewers

KB: Karla Ballman, Macalester; LB: Lynne Baur, Carleton; DB: David Bressoud, Macalester; LC: Laura Chihara, St. Olaf; CEC: Clifton E. Corzatt, St. Olaf; TH: Tom Halverson, Macalester; MK: Michael Kahn, St. Olaf; SK: Steve Kennedy, Carleton; RM: Richard Molnar, Macalester; JO: Jeff Ondich, Carleton; JS: John Schue, Macalester; LAS: Lynn Arthur Steen, St. Olaf; MW: Martha Wallace, St. Olaf.

THE AUTHORS

GEORGE COBB, Robert L. Rooke Professor of Statistics at Mount Holyoke College, was an undergraduate Russian major (A.B. Dartmouth) before turning to statistics in graduate school (M.S. Medical College of Virginia, Ph.D. Harvard). He currently chairs the Joint Committee on Undergraduate Statistics of the MAA and the American Statistical Association. Cobb maintains a precarious professional existence on the cusp between the scholarly and the sophomoric, where his recent efforts include "Three Ways to Gum Up a Statistics Course," a seminal paper in the emerging science of Confectionary Ballistics (a.k.a. launching Gummi Bears), and "Brains, Slime, and Elasticities," which uses Elmer's glue to cement a long-suspected connection between econometricians and paleontologist Stephen Jay Gould.

DAVID MOORE was trained in mathematics (A.B. Princeton, Ph.D. Cornell). His Cornell advisor, the late Jack Kiefer, told him the thesis in mathematical statistics was fine but "There's a lot of statistics every sociologist knows that you don't know." A few years in the Purdue Statistics Department began both to fill the gaps and change attitudes in ways that he hopes this paper communicates. He is now Shanti S. Gupta Distinguished Professor at Purdue and president-elect of the American Statistical Association. Once out of office, he hopes to return to such real-life enjoyments as traveling with his wife in France and Italy, gardening, and teaching Sunday School.

MARCO ABATE was born in Milan in 1962, got his Ph.D. from Scuola Normale Superiore, Pisa in 1988, and now is full professor of Geometry at the University of Ancona (placing him among the ten youngest Italian full professors). His interests include geometrical function theory, holomorphic dynamical systems, complex differential geometry, and traveling around the world. In his spare time, he writes comics, sometimes even about mathematics.

ERIC BACH was born in Chicago and grew up in Austin, Texas. He received his B.A. (in mathematics) from Michigan in 1974, and his Ph.D. (in computer science) from Berkeley in 1984, under the direction of Manuel Blum. Since then he has been on the faculty of the University of Wisconsin in Madison. He is a past recipient of the ACM Distinguished Dissertation Award and the Presidential Young Investigator Award. His research interests include computational number theory, analysis of algorithms, cryptography, complexity theory, and six-string automata.

THOMAS FORSTER was born in a well-known haunted house in Cambridge, U.K. in 1948. He went to university intending to study history but in Quine's *From a Logical Point of View* read the article whose 60th anniversary he is here commemorating and decided to study instead philosophy and later music and mathematics, completing a Ph.D. in set theory under Adrian Mathias and Maurice Boffa in 1976. After many years in New Zealand (including another false start in neurophysiology) he reluctantly returned to the city of his birth, with his wife and two stepdaughters, where he is now fellow and director of studies in mathematics and computer science in Hughes Hall, the oldest and poorest of the graduate colleges at the University of Cambridge.

MICHAŁ MISIUREWICZ was born in Warsaw, Poland in 1948. He lived there for the first 42 years of his life, with small breaks for doing math at other places. He received his Ph.D. from Warsaw University in 1974. He moved to the New World in 1990, first visiting Northwestern University, then Princeton University, and finally solving his two body problem by settling down as a professor of mathematics at Indiana University-Purdue University Indianapolis. He likes all kinds of dynamical systems, but usually prefers one-dimensional ones. He is fond of his tiny little points in the Mandelbrot set.

ANDREW GRANVILLE did his undergraduate education in England, his postgraduate education in Canada, and has subsequently been a professor in the U.S. This much traveling and confusion is probably what led to the mistake in the proof in the first version of this paper, but now that he's settled down in Georgia then so maybe is the proof!

VICTOR ADAMCHIK is a mathematician with Wolfram Research, Inc. the developers of *Mathematica*. He received his Ph.D. from the Byelorussian State University (Minsk, former USSR). His research areas in mathematics include theory of functions, integral transforms, asymptotic expansions, differential equations, and symbolic algorithms of summation and integration.

STAN WAGON is a professor of mathematics at Macalester College. He has long been intrigued by the way *Mathematica*'s power and breadth allow us to visualize mathematical constructions that were once considered extremely abstract, and has written several books on that theme. His favorite example along these lines is a computer visualization of a variation of the Banach–Tarski paradox.

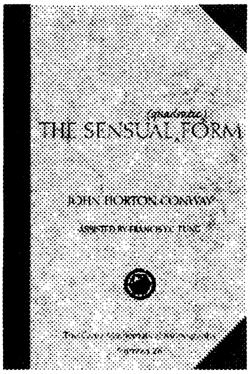
FRANCIS EDWARD SU earned his Ph.D. from Harvard in 1995 studying random walks with Persi Diaconis, but he has been a fan of the Brouwer Theorem since taking a Moore method course from Mike Starbird at UT-Austin. He is an assistant professor at Harvey Mudd College and an MAA Project NExT fellow. His research interests include random walks on homogeneous spaces, continued fractions, and fixed point theory. Rumor has it that he does classroom demos of Brouwer's Theorem with human subjects.

ZDZISŁAW POGODA studied at the Jagiellonian University in Kraków and obtained his doctorate in 1982. He works at the Institute of Mathematics of the Jagiellonian University. His main scientific interests are differential geometry and its applications, the history of mathematics (with emphasis on the 19th and early 20th centuries), and popularization of mathematics. He devotes a great deal of time to work with young people. In 1995 he, and his friend and frequent coauthor Krzysztof Ciesielski, obtained the Great Award of the Polish Mathematical Society for the popularization of mathematics. He and Krzysztof Ciesielski have written two popular-scientific books that explore various areas of mathematics.

LESZEK M. SOKOŁOWSKI studied at the Jagiellonian University in Kraków and obtained his doctorate in 1975. Since then he has been working at the Astronomical Observatory of that university. His present position is that of an associate professor of gravitational physics. His field of research includes relativistic cosmology, modern Kaluza-Klein theories, black-hole physics, and gravitational waves. He is also interested in the philosophical problems of physics.

BILL DUNBAR had a double concentration in math and Russian studies as an undergraduate at Brown University, and received his Ph.D. in mathematics from Princeton University in 1981, with a thesis written under the influence of Bill Thurston. He now teaches math and statistics at Simon's Rock College of Bard, where the students have all left high school early to start college. He has published a few papers on 3-dimensional orbifolds, and is now trying to see if he can sell crystallographers on the idea of using Euclidean orbifolds as a natural domain for electron density maps.

ANDREW BREMNER was raised in Lancashire and Yorkshire, Oxford and Cambridge, with a degree under the supervision of Swinnerton-Dyer. He taught at Cambridge between 1975 and 1984, before leaving for the U.S. South West and Arizona State University. Research interest is Diophantine equations. Non-research interests are varied, but include canyon, mountain, and desert hiking. Solo crossings of the Sahara and of Tibet are achievements in travel.



(quadratic)

The Sensual Form

Series: Carus Mathematical Monographs

John Horton Conway
Assisted by Francis Y. C. Fung

John Horton Conway's unique approach to quadratic forms was the subject of the Hedrick Lectures given by him in August 1991 at the Joint Meetings of the American Mathematical Society and the Mathematical Association of America in Orono, Maine. Professor Conway provides the following overview of those lectures:

"I have been interested in quadratic forms for many years, but keep on discovering new and simple ways to understand them. The "topograph" of the First Lecture makes the entire theory of binary quadratic forms so easy that we no longer need to think or prove theorems about these forms—just look! In some sense the experts knew something like this picture—but why did they use it only in the analytic theory, rather than right from the start?

Since sight and hearing were not involved, I took as the theme of the lectures the idea that one should try to appreciate quadratic forms with all one's senses, and so arose the title "The Sensual Form" for my Hedrick Lectures, and also the topics for the first two of them.

I could not settle on a single topic for the third of these lectures, even when I came to give it. So, in the end, I split it into two half-hour talks, one on the shape of the Voronoi cell of a lattice, and one on the Hasse-Minkowski theory. In this book, each of the these has become a fully-fledged lecture.

The book should not be thought of as a serious textbook on the theory of quadratic forms—it consists rather of a number of essays on particular aspects of quadratic forms that have interested me. The lectures are self-contained and will be accessible to the generally informed reader who has no particular background in quadratic form theory. The minor exceptions should not interrupt the flow of ideas. The Afterthoughts to the Lectures contain discussions of related matters that occasionally presuppose greater knowledge.

Since so much of the treatment is new to this book, it may not be easy to circumvent one's difficulties by reference to standard texts. I hope the work pays off, and that even the experts in quadratic forms will find some new enlightenment here." —John Horton Conway

Contents:
First Lecture: Can you see the values of $3x^2 + 6xy - 5y^2$?; Afterthoughts: $\mathbf{PSL_2(Z)}$ and Farey Fractions;
Second Lecture: Can You Hear the Shape of a Lattice?; Afterthoughts: Kneser's Gluing Method: Unimodular Lattices;
Third Lecture: ...and Can You Feel its Form?; Afterthoughts: Feeling the Form of a Four-Dimensional Lattice;
Fourth Lecture: The Primary Fragrances; Afterthoughts: More about the Invariants: The p-Adic Numbers;
Postscript: A Taste of Number Theory.

Catalog Code: CAM-26/JR
225 pp., Hardbound, 1997
ISBN 0-88385-030-3
List: $32.95 MAA Member: $25.95

Phone in Your Order Now! ☎ 1-800-331-1622

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows **USA orders (shipped via UPS):** $2 95 for the first book, and $1.00 for each additional book **Canadian orders:** $4.50 for the first book and $1 50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route We do not ship via UPS into Canada unless the customer specially requests this service Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders** $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7 00 per book Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse Credit Card orders are accepted for all customers

Name ____________________

Address ____________________

City ______________ State _____ Zip ________

Phone ____________________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	CAM-26/JR	______	______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling ________

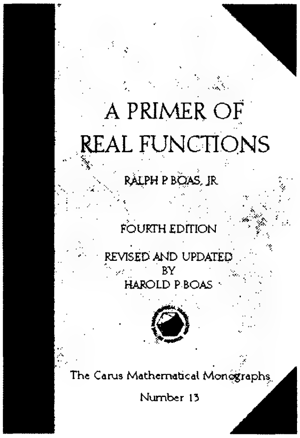
TOTAL ________

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________________ Expires ___/___

Signature ____________________

THE MATHEMATICAL ASSOCIATION OF AMERICA



A Primer of Real Functions

by Ralph P. Boas

Revised and updated by Harold P. Boas

Series: Carus Mathematical Monograph

This is a revised, updated and augmented edition of a classic Carus monograph (a bestseller for over 25 years) on the theory of functions of a real variable. Earlier editions of this classic Carus Monograph covered sets, metric spaces, continuous functions, and differentiable functions. The fourth edition adds sections on measurable sets and functions, the Lebesgue and Stieltjes integrals, and applications. The book is accessible to readers with some mathematical sophistication and a background in calculus. It is suitable either for self-study or for supplemental reading in a course on advanced calculus or real analysis.

Not intended as a systematic treatise, this book has more the character of a sequence of lectures on a variety of topics connected with real functions. Many of these topics are not commonly encountered in undergraduate textbooks: for example, the existence of continuous everywhere-oscillating functions (via the Baire category theorem); two functions having equal derivatives, yet not differing by a constant; application of Stieltjes integration to the speed of convergence of infinite series.

Table of Contents:

I. Sets: Sets of real numbers, Countable and uncountable sets, Metric spaces, Open and closed sets, Dense and nowhere dense sets, Compactness, Convergence and completeness, Nested sets and Baire's theorem, Some applications of Baire's theorem, Sets of measure zero. II. Functions: Functions, Continuous functions, Properties of continuous functions, Upper and lower limits, Sequences of functions, Uniform convergence, Pointwise limits of continuous functions, Approximations to continuous functions, Linear functions, Derivatives, Monotonic functions, Convex functions, Infinitely differentiable functions. III. Integration: Lebesgue measure, Measurable functions, Definition of the Lebesgue integral, Properties of Lebesgue integrals, Application of the Lebesgue integral, Stieltjes integrals, Applications of the Stieltjes integral, Partial sums of infinite series.

Catalog Code: CAM-13R/JR
262 pp., Hardcover, 1996
ISBN 0-88385-029-X
List: $35.95 MAA Member: $24.95

Phone in Your Order Now! ☎ 1-800-331-1622

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders**: $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ______________________

Address ______________________

City ____________ State _____ Zip ________

Phone ______________________

QTY.	CATALOG CODE	PRICE	AMOUNT
_______	CAM-13R/JR	_______	_______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling _______

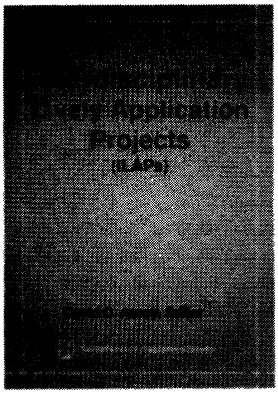
TOTAL _______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

THE MATHEMATICAL ASSOCIATION OF AMERICA



Interdisciplinary Lively Application Projects (ILAPs)

Series: Classroom Resource Materials

David C. Arney

Interdisciplinary Lively Application Projects (ILAPs) are small group projects developed through the cooperation of faculty from mathematics and partner disciplines. These projects will provide teachers with material that can help their students understand mathematical concepts, develop strong mathematical skills and motivate them towards an interest in future subjects accessible through the study of mathematics. It is an important step towards helping students acquire a broad, interdisciplinary outlook towards problem solving.

The ILAPs provide supplemental classroom resource materials in the form of eight project handouts that you can use as student homework assignments. They require students to use scientific and quantitative reasoning, mathematical modeling, symbolic manipulation skills, and computational tools to solve and analyze scenarios, issues, and questions involving one or more disciplines. Sample solutions to the problems, background material, notes to the instructor and a student writing guide are also included.

The prerequisite skills for the eight projects presented in the book range from freshmen-level algebra, trigonometry, and precalculus; through calculus, elementary and intermediate differential equations, and discrete mathematics to advanced calculus and partial differential equations. The partner disciplines included in the projects are: mechanics, physics, chemistry, engineering, geography, topography, and exercise physiology. You can use the projects as a supplement to a textbook in any of the following undergraduate areas: precalculus, calculus, linear algebra, differential equations, discrete mathematics, mathematical modeling, advanced calculus, partial differential equations, and numerical computing.

The book also contains several supporting articles that describe uses for these projects.

Contents: **ILAPs:** Getting Fit with Mathematics; Decked Out; Parachute Panic; Flying with Differential Equations; Planning a Backpacking Trip to Pikes Peak; SMOG in Los Angeles Basin, Structural Mechanics — Beams and Bridges; Contaminant Transport. **Articles:** Technical Report Format and Writing Guide; Project INTERMATH: An Interdisciplinary Approach to Cultural Change; ILAP Products: Authoring, Testing and Editing; Interdisciplinary and Integrated Curriculum Models; Interdisciplinary Communication and Understanding; Interdisciplinary Projects at West Point.

Catalog Code: ILAP/JR
206 pp., Paperbound, 1997
ISBN 0-88385-706-5
List: $27.50 MAA Member: $ 22.00

Phone in Your Order Now! ☎ 1-800-331-1622

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders:** $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ______________________

Address ______________________

City ____________ State ____ Zip ______

Phone ______________________

Qty.	Catalog Code	Price	Amount
______	ILAP/JR	______	______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling ______

TOTAL ______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

UNITED STATES POSTAL SERVICE

Statement of Ownership, Management, and Circulation

(Required by 39 USC 3685)

1 Publication Title	2 Publication Number	3 Filing Date
The American Mathematical Monthly	0 0 0 2 – 9 8 9 0	September 9, 1997
4 Issue Frequency	**5 Number of Issues Published Annually**	**6 Annual Subscription Price**
Monthly, except bimonthly June/July, Aug./Sep	10	$38

7 Complete Mailing Address of Known Office of Publication *(Not printer) (Street city county, state, and ZIP+4)*
MAA, 1529 Eighteenth St , NW, Washington, DC 20036-1385

Contact Person: Harry Waldman
Telephone: 202-387-5200

8 Complete Mailing Address of Headquarters or General Business Office of Publisher *(Not printer)*
Same

9 Full Names and Complete Mailing Addresses of Publisher Editor and Managing Editor *(Do not leave blank)*

Publisher *(Name and complete mailing address)*
The Mathematical Association of America, 1529 Eighteenth St., NW, Washington, DC 20036-1385

Editor *(Name and complete mailing address)*
Roger Horn, University of Utah, Salt Lake City, Utah 84112

Managing Editor *(Name and complete mailing address)*
Harry Waldman, MAA, 1529 18th St., NW, Washington, DC 20036-1385

10 Owner *(Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm give its name and address as well as those of each individual owner If the publication is published by a nonprofit organization, give its name and address.)*

Full Name	Complete Mailing Address
The Mathematical Association of America	1529 Eighteenth St., NW, Washington, DC 20036

11 Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds Mortgages, or Other Securities If none check box → ☒ None

Full Name	Complete Mailing Address

12 Tax Status *(For completion by nonprofit organizations authorized to mail at special rates) (Check one)*
The purpose function and nonprofit status of this organization and the exempt status for federal income tax purposes
☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months *(Publisher must submit explanation of change with this statement)*

PS Form **3526**, September 1995 *(See Instructions on Reverse)*

13 Publication Title	14 Issue Date for Circulation Data Below	
The American Mathematical Monthly	August–September 1997	
15 Extent and Nature of Circulation	**Average No. Copies Each Issue During Preceding 12 Months**	**Actual No. Copies of Single Issue Published Nearest to Filing Date**
a Total Number of Copies *(Net press run)*	20,500	20,350
b Paid and/or Requested Circulation (1) Sales Through Dealers and Carriers, Street Vendors, and Counter Sales *(Not mailed)*	0	0
b Paid and/or Requested Circulation (2) Paid or Requested Mail Subscriptions *(Include advertiser's proof copies and exchange copies)*	17,000	15,200
c Total Paid and/or Requested Circulation *(Sum of 15b(1) and 15b(2))*	17,000	15,200
d Free Distribution by Mail *(Samples, complimentary and other free)*	1,750	1,750
e Free Distribution Outside the Mail *(Carriers or other means)*	0	0
f Total Free Distribution *(Sum of 15d and 15e)*	1,750	1,750
g Total Distribution *(Sum of 15c and 15f)*	18,750	16,950
h Copies not Distributed (1) Office Use Leftovers Spoiled	1,750	3,400
h Copies not Distributed (2) Returns from News Agents	0	0
i Total *(Sum of 15g 15h(1) and 15h(2))*	20,500	20,350
Percent Paid and/or Requested Circulation *(15c / 15g x 100)*	91%	90%

16 Publication of Statement of Ownership
☒ Publication required. Will be printed in the November 1997 issue of this publication
☐ Publication not required.

17 Signature and Title of Editor, Publisher Business Manager or Owner
Donald J. Albers

Date 9/9/97

I certify that all information furnished on this form is true and complete I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including multiple damages and civil penalties)

Instructions to Publishers

1 Complete and file one copy of this form with your postmaster annually on or before October 1 Keep a copy of the completed form for your records

2 In cases where the stockholder or security holder is a trustee, include in items 10 and 11 the name of the person or corporation for whom the trustee is acting Also include the names and addresses of individuals who are stockholders who own or hold 1 percent or more of the total amount of bonds, mortgages, or other securities of the publishing corporation In item 11, if none, check the box Use blank sheets if more space is required

3 Be sure to furnish all circulation information called for in item 15 Free circulation must be shown in items 15d, e, and f

4 If the publication had second-class authorization as a general or requester publication, this Statement of Ownership, Management, and Circulation must be published: it must be printed in any issue in October or, if the publication is not published during October, the first issue printed after October

5 In item 16, indicate the date of the issue in which this Statement of Ownership will be published

6 Item 17 must be signed

Failure to file or publish a statement of ownership may lead to suspension of second-class authorization

PS Form **3526**, September 1995 *(Reverse)*

INDEX TO VOLUME 104, 1997
THE AMERICAN MATHEMATICAL MONTHLY

TITLE INDEX

Author Index

Problems Proposed

Problems Solved

SOLUTIONS

Numbers in boldface refer to problems; those in lightface to pages.

THANKS

The MONTHLY expresses its appreciation to the following people for their help in refereeing during the past year. We could not function successfully without such people and their hard work.

Arnold M. Adelberg
Don Albers
George E. Andrews
Jeff Angell
Kenneth I. Appel
Tom Archibald
Michael Aschbacher
Richard Askey
David Auckly
Sheldon Axler
Laszlo Babai
Duane Bailey
Joseph A. Ball
Thomas F. Banchoff
William Barker
Robert Bartle
Tamer Basar
Estelle Basor
Mike Beals
William Becker
Adi Ben-Israel
Elwyn Berlekamp
Rajendra Bhatia
Ken Binmore
Tibor Bisztriczky
William Blair
Andreas Blass
Harold Boas
Kenneth P. Bogart
Edward M. Bolger
Jonathan M. Borwein
Peter Borwein
Doug Bowman
David M. Bradley
Fred Brauer
Andrew Bremner
David Bressoud
Jason Brown
Morton Brown

Robert Brown
Richard A. Brualdi
Andy Bruckner
Barry W. Brunson
Robert Burckel
David M. Burton
Sam Buss
Ron Calinger
David H. Carlson
Glenn Chappell
Tim Chow
Dietmar Cieslik
Erhan Cinlar
Colin Clark
Mark Copper
Robert M. Corless
Lee Corry
Tom Cover
Carl Cowen
Charles Cullen
Al Cuoco
Jim Cushing
John D'Angelo
Richard B. Darst
Robert Dawson
Carl de Boor
Francois de Gandt
Jesus A. deLoera
James Demmel
Robert L. Devaney
Persi Diaconis
Harold Diamond
Randall Dougherty
Tommy Dreyfus
Richard M. Dudley
Dwight Duffus
Alan S. Edelman
Saber N. Elaydi
Susanna Epp
Herman Erlichson
James E. Falk
Ralph J. Faudree, Jr.
Temple Harold Fay
Thomas S. Ferguson
Michael Filaseta
Jarñes Fill
Carl FitzGerald
Mary Flahive
Harley Flanders
Sergey Fomin
James Foran
E. A. D. Foster
David Fowler
Michael Frame
Craig Fraser
Roland W. Freund
Steve Gadbois
Fred Galvin
Ted Gamelin
Murray Gerstenhaber
Ira Gessel
Etienne Ghys
Len Gillman
Geoff Goodson
Russell A. Gordon
Daniel Goroff
Judith V. Grabiner
Ronald L. Graham
Hardy Grant
Andrew J. Granville
I. Grattan-Guinness
Jerrold R. Griggs
Nathaniel Grossman
Branko Gruenbaum
Gary Gruenhage
John M. Guckenheimer
Niccolo Guicciardini
Eugene Gutkin
Mowaffaq Hajja
Philip Hanlon
Kathy Hann
Randall Holmes
Lars Holst
Richard Holzsager
Fredric T. Howard
Thomas C. Hull
Thomas Jech
Joe Jerome
Charles R. Johnson
Keith Johnson
Elgin Johnston
Hugo D. Junghenn
James Kaput
Irving J. Katz
Kiran S. Kedlaya
Tim Kelly
Emelie Kenney
Darrell Kent
Henry A. Kierstead
Jeremy Kilpatrick
Clark Kimberling
Victor Klee
Dennis Kletzing
Ron Knill
Donald E. Knuth
Roman Kossak
Roger Kraft
Jeffrey C. Lagarias
T.Y. Lam
Detlef Laugwitz
Greg Lawlor
Frank T. Leighton
Steven Leon
Richard Lesh
Gail Letzter
Simon Levin
Elliott Lieb
W. Brent Lindquist
Peter Loeb
Calvin T. Long
Franklin Lowenthal
Daniel H. Luecking
Tom MacGregor
Kirill C. H. Mackenzie
George Mackiw
Hosam M. Mahmoud
Olvi L. Mangasarian
Marvin Marcus
Ellen Maycock Parker
Bernard J. McCabe
Cathy McGeoch
Russell Merris
Carl D. Meyer
T.S. Michael
Arnie Miller
John Milnor
Ray Mines
Rennie Mirollo
John A. Mitchem
Hugh L. Montgomery
Frank Morgan
S. Brent Morris
Michael F. Moses
Paul S. Muhly
Mel Nathanson
R.D. Neidinger
Donald J. Newman
Morris Newman
Yves Nievergelt
Peter Nyikos
James E. Nymann
George O'Brien
Ellen O'Keefe
Robert E. O'Malley, Jr.
Robert H. Oehmke
Andrew P. Ogg
Adrian Oldknow
Robert F. Olin
Vadim Olshevsky
Jeffrey R. Ondich
Ken Ono
Harold R. Parks
Beresford N. Parlett
Brian J. Parshall
Seymour Parter

Michael Perlman
Washek Pfeffer
Jim Pitman
Henry Pollak
Andy Pollington
Bjorn Poonen
Horia Pop
Stephen Portnoy
Carey Priebe
James G. Propp
Robert Propp
Stanley Rabinowitz
Tom Ransford
Peter Renz
Norman J. Richert
Fred Richmon
David Robbins
Donald Robinson
E. Arthur Robinson
Eleanor Robson
Andrew M. Rockett
Leiba Rodman
Yongwu Rong
Kenneth A. Ross
Gian-Carlo Rota
Uriel Rothblum
Bruce Rothschild
Cecil C. Rousseau
J. Rovnyak
Mark Rubenstein
Mary Ellen Rudin
Jeff Sachs
Lee Sallows
Paul Samuelson
Thomas R. Scavo
Frank W. Schmidt
Larry Schumaker
Eugene Schuster
John Selden
Peter Semrl
Marjorie Senechal
Bryan Shader
Glenn Shafer
Jeffrey O. Shallit
Joel Shapiro
Louis W. Shapiro
Lawrence A. Shepp
Ted Shifrin
Don Shimamoto
Steven Shreve
Wilfried Sieg
Alice Silverberg
Rodica E. Simion
Stephen G. Simpson
Ya. G. Sinai
Joel A. Smoller
Wasin So
H. Mete Soner
Roy Sorenson
Larry Sowder
Joel H. Spencer
Eugene Spiegel
Richard P. Stanley
J. Michael Steele
Juris Steprans
John Stillwell
Arthur H. Stone
Philip Straffin, Jr.
Steven Strogatz
Keith Stroyan
Ted Suffridge
Howard Swann
Michel Talagrand
Elliot Tanis
James J. Tattersall
Jeremy Teitelbaum
Walter Thirring
John B. Thoo
John Todd
Gabor Toth
Lloyd N. Trefethen
Hale Trotter
William T. Trotter
Daniel J. Troy
Thomas W. Tucker
Tom Tymoczko
Jerry Uhl
Frank D. Uhlig
Charles Van Loan
Cris Van Wyck
Charles L. Vanden Eynden
Ilan Vardi
Richard Varga
Jerry Vaughan
Stan Wagon
Samuel S. Wagstaff, Jr.
Jim Walsh
Peter L. Waterman
Stephen Watson
Frank Wattenberg
Jeffrey Weeks
Marysia T. Weiss
John Wermer
Robert Wheeler
Ben White
Derek T. Whiteside
Sylvia Wiegand
John Wierman
Robert F. Williams
Curtis Wilson
Harald Wimmer
H. H. Wu
Norman Zadeh
Lawrence Zalcman
Doron Zeilberger
William Zwicker

THE AMERICAN MATHEMATICAL
MONTHLY

Volume 104, Number 10 December 1997

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and Notes should be sent to the Editor:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send your email address and 3 copies of the complete manuscript (including all figures with captions and lettering), typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper.

Letters to the Editor on any topic are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

See the MONTHLY section of MAA Online for current information such as contents of issues, descriptive summaries of forthcoming articles, tips for authors, and preparation of manuscripts in TEX:

http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all problems/solutions material, typewritten on only one side of the paper.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Mr. JOE WATSON, Advertising Manager

Change of address, missing issues inquiries, and other subscription correspondence:
MAA Service Center
maahq@maa.org

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Ramanujan's Association with Radicals in India

Bruce C. Berndt, Heng Huat Chan, and Liang-Cheng Zhang

In memory of Ramanujan on the

$$\left(32\left(\frac{146410001}{48400}\right)^3 - 6\left(\frac{146410001}{48400}\right) + \sqrt{\left(32\left(\frac{146410001}{48400}\right)^3 - 6\left(\frac{146410001}{48400}\right)\right)^2 - 1}\right)^{1/6} \text{th}$$

anniversary of his birth.

Ramanujan (sometimes under the alternative spelling Ramanujam) submitted 58 problems to the *Journal of the Indian Mathematical Society*. Approximately ten of them involve equalities between radicals. For example [**16**, p. 334],

$$\left(\sqrt[5]{\tfrac{32}{5}} - \sqrt[5]{\tfrac{27}{5}}\right)^{1/3} = \sqrt[5]{\tfrac{1}{25}} + \sqrt[5]{\tfrac{3}{25}} - \sqrt[5]{\tfrac{9}{25}}.$$

Establishing equalities among exotic radicals was very common in Ramanujan's day, especially in Great Britain and its empire. For example, see the then popular texts by H. S. Hall and S. R. Knight [**12**, Chap. 8] and G. Chrystal [**10**, Chap. 11], the latter being well-known to Ramanujan. Was Ramanujan's keen interest in radical equalities merely a consequence of their popularity in his time, or were there other reasons? The answer can be found in his notebooks [**15**] and in one of his most important papers [**14**], [**16**, pp. 23–39].

Scattered among the pages in Ramanujan's first notebook are the values of 107 class invariants, or polynomials satisfied by them. As we shall see, these invariants frequently take the shapes of interesting radicals, and often to put the radical expressions in their most attractive forms, difficult radical equalities need to be established. So that we may define Ramanujan's class invariants, set

$$\chi(q) = \prod_{k=0}^{\infty}(1 + q^{2k+1}).$$

For any positive rational number n, set

$$q = \exp(-\pi\sqrt{n}),$$

and define the two *class invariants* G_n and g_n by

$$G_n := 2^{-1/4}q^{-1/24}\chi(q) \qquad \text{and} \qquad g_n := 2^{-1/4}q^{-1/24}\chi(-q). \tag{1}$$

As we shall see, we are able to calculate G_n for certain odd values of n and g_n for certain even values of n.

At the beginning of the twentieth century, these invariants were extensively studied by H. Weber [**21**], who used the notations $G_n =: 2^{-1/4}\mathfrak{f}(\sqrt{-n})$ and $g_n =: 2^{-1/4}\mathfrak{f}_1(\sqrt{-n})$. Weber [**21**] proved that G_n and g_n are algebraic. In fact, G_n, $2^{-1/12}G_n$, and $2^{-1/4}G_n$ are units in some algebraic number field according as $n \equiv 1 \pmod 4$, $n \equiv 3 \pmod 8$, and $n \equiv 7 \pmod 8$, respectively. If $n \equiv 2 \pmod 4$, then g_n is a unit. Weber's study of G_n and g_n was motivated by the construction of

the *Hilbert class field* H_n, the maximal unramified abelian extension of the imaginary quadratic field $K_n := \mathbb{Q}(\sqrt{-n})$. It can be shown that $H_n = K_n(j(\omega_n))$, where

$$\omega_n = \begin{cases} \sqrt{-n}, & \text{if } n \equiv 1 \pmod 4, \\ \dfrac{3+\sqrt{-n}}{2}, & \text{if } n \equiv 3 \pmod 4, \end{cases}$$

and j is the famous *modular j-invariant*, so-called because it is invariant under transformations from the modular group. Weber **[21]** asserted that certain small powers of G_n and g_n can be used to replace $j(\omega_n)$ as generators of H_n over K_n. Perhaps for these reasons, Weber called G_n and g_n *class invariants*, and computed a total of 105 class invariants or the monic, irreducible polynomials satisfied by them. The excellent text of D. A. Cox **[11]** provides an accessible account of Weber's work on invariants.

Before proceeding further, we give some examples that Ramanujan calculated:

$$G_5 = \left(\frac{1+\sqrt{5}}{2}\right)^{1/4}, \tag{2}$$

$$G_{17} = \sqrt{\frac{5+\sqrt{17}}{8}} + \sqrt{\frac{\sqrt{17}-3}{8}},$$

$$G_{69} = \left(\frac{5+\sqrt{23}}{\sqrt{2}}\right)^{1/12}\left(\frac{3\sqrt{3}+\sqrt{23}}{2}\right)^{1/8}\left(\sqrt{\frac{6+3\sqrt{3}}{4}} + \sqrt{\frac{2+3\sqrt{3}}{4}}\right)^{1/2}, \tag{3}$$

and

$$G_{1353} = \left(\frac{3+\sqrt{11}}{\sqrt{2}}\right)^{1/4}\left(\frac{5+3\sqrt{3}}{\sqrt{2}}\right)^{1/4}\left(\frac{11+\sqrt{123}}{\sqrt{2}}\right)^{1/4}\left(\frac{6817+321\sqrt{451}}{\sqrt{2}}\right)^{1/12}$$
$$\times\left(\sqrt{\frac{17+3\sqrt{33}}{8}} + \sqrt{\frac{25+3\sqrt{33}}{8}}\right)^{1/2}$$
$$\times\left(\sqrt{\frac{561+99\sqrt{33}}{8}} + \sqrt{\frac{569+99\sqrt{33}}{8}}\right)^{1/2}.$$

The value of G_{69} was only recently verified for the first time by the authors **[7]**. In our calculation of G_{69}, we used the equality

$$\left(188+108\sqrt{3}+\sqrt{(188+108\sqrt{3})^2-1}\right)^{1/6} = \sqrt{\frac{6+3\sqrt{3}}{4}} + \sqrt{\frac{2+3\sqrt{3}}{4}},$$

which is the special case, $a = (4+3\sqrt{3})/4$, of the more general equality

$$\left(32a^3-6a+\sqrt{(32a^3-6a)^2-1}\right)^{1/6} = \sqrt{a+\tfrac{1}{2}} + \sqrt{a-\tfrac{1}{2}},$$

which we also used in the dedication at the beginning of this paper and in calculations of further class invariants.

The value of G_{1353} was communicated by Ramanujan **[16**, p. xxix, eq. (23)], **[8**, p. 62] in his second letter, dated 27 February 1913, to G. H. Hardy and was first established (unrigorously) by G. N. Watson **[18]**. In a letter of 1 October 1930 to

B. M. Wilson [**8**, pp. 237, 238], Watson confided, "... but 23 which deals with the singular modulus associated with 1353 is included; I was pleased at getting this out, because the bulk of the singular moduli in the Notebooks can be obtained in the same way ... You will be interested to hear how Ramanujan got no. 23, particularly when you look at the length of the answer. I am absolutely convinced that he guessed it." (Calculating a singular modulus, which we do not define here, is equivalent to calculating a class invariant.) The reader is undoubtedly astonished to learn that Ramanujan first "guessed" his formula for G_{1353}. We do not agree with Watson! We think that Watson's proof, which is not rigorous, could not have been given without his knowing the formula in advance. The first rigorous proof was given recently by Chan [**9**].

On pages 294–299 in his second notebook [**15**], Ramanujan gave a table of values for 77 class invariants, three of which are not found in the first notebook. Since the second notebook is an enlarged revision of the first, it is unclear why Ramanujan failed to record 33 class invariants that he offered in the first notebook. Four further results are found in scattered places in the second notebook. After arriving in Cambridge, Ramanujan learned of Weber's work [**21**], and so when he wrote his paper [**14**], [**16**, pp. 23–39], the table of 46 class invariants that he included did not contain any that are found in Weber's book [**21**]. Except for G_{325} and G_{363}, all of the remaining values are found in Ramanujan's notebooks. To the best of our reckoning, Ramanujan calculated a total of 116 class invariants, or monic, irreducible polynomials satisfied by them.

Why did Ramanujan calculate such a large number of class invariants? Ramanujan did not share Weber's interest in generating Hilbert class fields, but he did have applications. First, as the title of his paper [14] indicates, Ramanujan used class invariants to find excellent approximations to π. For example, from (1) and (3), we find that

$$\pi \approx \frac{24}{\sqrt{69}}\left(\log G_{69} + \tfrac{1}{4}\log 2\right) = 3.1415926536032\ldots,$$

which agrees with the value of π through nine decimal places.

Second, Ramanujan used class invariants to determine explicitly particular values of the theta function $\varphi(q)$ defined by

$$\varphi(q) := \sum_{k=-\infty}^{\infty} q^{k^2}.$$

For example, Ramanujan probably used his value of G_{49} to show that [**3**]

$$\frac{\varphi^2(e^{-7\pi})}{\varphi^2(e^{-\pi})} = \frac{\sqrt{13+\sqrt{7}} + \sqrt{7+3\sqrt{7}}}{14}(28)^{1/8}. \tag{4}$$

The value

$$\varphi(e^{-\pi}) = \frac{\pi^{1/4}}{\Gamma\left(\frac{3}{4}\right)}$$

is well known [**1**, p. 103], and so (4) provides an explicit evaluation for $\varphi(e^{-7\pi})$.

The theta function φ is intimately connected with elliptic functions and integrals. An elliptic function is a function of a complex variable with two linearly independent periods, in contrast to the familiar trigonometric functions, which have just one linearly independent period. The complete elliptic integral of the

first kind associated with the *modulus* k, $0 < k < 1$, is defined by

$$K := K(k) := \int_0^{\pi/2} \frac{d\theta}{\sqrt{1 - k^2 \sin^2\theta}}. \tag{5}$$

The complementary modulus k' is defined by $k' = \sqrt{1 - k^2}$; set $K' = K(k')$. If $q = \exp(-\pi K'/K)$, then one of the central theorems in the theory of elliptic functions asserts that

$$\varphi^2(q) = \frac{2}{\pi} \int_0^{\pi/2} \frac{d\theta}{\sqrt{1 - k^2 \sin^2\theta}} = \frac{2}{\pi} K(k). \tag{6}$$

By (6), an explicit determination of $\varphi(q)$ for a certain value of q also yields an explicit value for $K(k)$.

A second classical theta function is the Dedekind eta-function $\eta(z)$ defined by

$$f(-q) := q^{-1/24}\eta(z) := \prod_{k=1}^{\infty} (1 - q^k) = \sum_{k=-\infty}^{\infty} (-1)^k q^{k(3k-1)/2}, \tag{7}$$

where $q = \exp(2\pi i z)$ and $|q| < 1$. The exponents $k(3k - 1)/2$ are called pentagonal numbers, and the second equality in (7) constitutes Euler's pentagonal number theorem. From (1) and (7), we easily see that

$$G_n = 2^{-1/4} q^{-1/24} \frac{f(q)}{f(-q^2)} \quad \text{and} \quad g_n = 2^{-1/4} q^{-1/24} \frac{f(-q)}{f(-q^2)}, \tag{8}$$

when $q = \exp(-\pi\sqrt{n})$. Ramanujan likely used his values of class invariants to calculate explicitly certain products of eta-functions in both his first and lost notebooks **[15]**, **[17]**. For example, he probably used the values of G_{225} and $G_{25/9}$ to prove that

$$e^{6\pi/5} \frac{f(-e^{-6\pi/5})}{f(-e^{-30\pi})} = \frac{a + b}{a - b}, \tag{9}$$

where $a = (60)^{1/4}$ and $b = 2 - \sqrt{3} + \sqrt{5}$.

Ramanujan also used class invariants to determine values of the celebrated Rogers-Ramanujan continued fraction $R(q)$, defined by

$$R(q) := \frac{q^{1/5}}{1} \; \frac{q}{+\,1} \; \frac{q^2}{+\,1} \; \frac{q^3}{+\,1} + \cdots, \quad |q| < 1.$$

The behavior of $R(q)$ for $|q| = 1$ is not completely understood, but if you have had a course in elementary number theory, perhaps you have shown that

$$R(1) = \frac{\sqrt{5} - 1}{2}.$$

Using the value of G_5, given in (2), we can show that

$$R^5(e^{-2\pi/\sqrt{5}}) = \sqrt{\left(\frac{5\sqrt{5} + 11}{2}\right)^2 + 1} - \frac{5\sqrt{5} + 11}{2}.$$

To offer another example [4], Ramanujan undoubtedly used (9) to show that

$$R(e^{-6\pi}) = \sqrt{c^2+1} - c,$$

where

$$2c := 1 + \frac{a+b}{a-b}\sqrt{5}.$$

Ramanujan calculated several further values of $R(q)$ in his lost notebook [**17**], and many of these can be found in the authors' paper [**6**].

In the remainder of the paper, we briefly describe some attempts and methods used to establish Ramanujan's class invariants.

In two papers [**19**], [**20**], Watson proved the 24 class invariants from Ramanujan's paper [**14**] that cannot be found in Ramanujan's second notebook. In the first [**19**], Watson devised an "empirical process" to calculate 14 of the 24 invariants, while in the second [**20**], he employed modular equations, which we define later in this paper, to prove 10 invariants. Watson [**18**] also used his empirical process to establish Ramanujan's value for G_{1353}. In the introduction to [**19**], Watson remarked, "It is intended to publish the calculations involved in the construction of the set $N+Q$ (the invariants appearing in both Ramanujan's paper [**14**] and the second notebook) as part of the commentary on the note-books by Dr. B. M. Wilson and myself." Although Watson and Wilson's efforts to edit Ramanujan's notebooks have been preserved in the library at Trinity College, Cambridge, Watson's calculations of these twenty-one invariants are not found there. The twenty-one values of n are: 65, 69, 77, 81, 117, 141, 145, 147, 153, 205, 213, 217, 265, 289, 301, 441, 445, 505, 553, 90, and 198. Watson wrote four further papers on the calculation of class invariants, and in those he verified three additional class invariants determined by Ramanujan, namely, those for $n = 81$, 147, and 289. Thus, after Watson's work, and up until recent times, 18 of Ramanujan's class invariants remained to be verified.

For five of the values, n is a multiple of 9, namely, $n = 117$, 153, 441, 90, and 198. The authors [**5**] found proofs for these values by using formulas relating G_{9n} with G_n and g_{9n} with g_n, which we established by using one of Ramanujan's modular equations of degree 3. All of the remaining 13 values are for G_n, $n = 65$, 69, 77, 141, 145, 205, 213, 217, 265, 301, 445, 505, and 553. Note that each value of n is the product of a small prime (3, 5, or 7) and a larger prime. Quite remarkably, the class number for each of the 13 imaginary quadratic fields $\mathbb{Q}(\sqrt{-n})$ equals 8. Moreover, there are precisely two classes per genus in each case. This is amazing! It is extremely unlikely that Ramanujan had any knowledge of imaginary quadratic fields, genus theory, or class numbers. However, Ramanujan must have recognized some arithmetical properties shared by these fields, although he would have expressed his ideas in a language very different from what we use today. How did Ramanujan calculate these 116 class invariants? He left no clues in his notebooks. Since Weber's methods were highly algebraic, it is very unlikely that Ramanujan journeyed along Weber's paths.

In his paper [**14**], Ramanujan used modular equations to calculate only a couple of simple invariants. This fact and the sentence, "The values of G_n and g_{2n} are got from the same modular equation." [**14**], [**16**, p. 25] are the only clues to his methods that Ramanujan provided for us. It would seem that if Ramanujan had employed another type of reasoning, he would have dropped some hint about it.

Having mentioned modular equations three times already in this paper, a definition of a modular equation is overdue.

With the elliptic integral K defined by (5), let K, K', L, and L' denote complete elliptic integrals of the first kind associated with the moduli k, k', l, and $l' := \sqrt{1-l^2}$, respectively, where $0 < k, l < 1$. Suppose that

$$n\frac{K'}{K} = \frac{L'}{L} \tag{10}$$

for some positive integer n. A relation between k and l induced by (10) is called a *modular equation of degree n*. In fact, modular equations are algebraic equations. A modulus can be expressed in terms of classical theta functions. Although we suppress all details, this fact, (10), and (6) are the primary ingredients needed to show that a modular equation can alternatively be expressed as an identity relating theta functions with argument q and theta functions with argument q^n.

As mentioned above, Watson **[20]** used modular equations to establish some of Ramanujan's invariants. We have been able to prove six of the remaining thirteen values for G_n, namely, for $n = 65, 69, 77, 141, 145$, and 213, by using modular equations of degrees p and q, where $n = pq$, but our approach is necessarily different from that of Watson. To prove the remaining seven invariants by employing modular equations, we would need modular equations of degrees 31, 41, 43, 53, 79, 89, and 101. Apparently, only for degree 31 did Ramanujan derive a modular equation, for he recorded no modular equations for the other six degrees in his notebooks. Some of the modular equations that we employed are very complicated, and so we had to use *Mathematica* to effect some of our calculations. In conclusion, it seems unlikely that Ramanujan used only modular equations in these elusive computations.

In order to prove the remaining class invariants of Ramanujan, we devised two methods **[7]**.

The first uses Kronecker's limit formula. In order to give a brief description of this formula, we need to define the Epstein zeta-function. Let $Q(u,v) := y^{-1}(u + vz)(u + v\bar{z})$, where $z = x + iy$ with $y > 0$. The Epstein zeta-function $\zeta_Q(s)$ is defined for $\sigma = \operatorname{Re} s > 1$ by

$$\zeta_Q(s) := \sum_{u,v} \{Q(u,v)\}^{-s},$$

where the sum is over all pairs of integers (u, v) except $(0, 0)$. It is well known that $\zeta_Q(s)$ can be analytically continued to the entire complex s-plane, where $\zeta_Q(s)$ is analytic except for a simple pole at $s = 1$. The Kronecker limit formula provides the constant term in the Laurent expansion about $s = 1$. This constant term involves the Dedekind eta-function, which we defined in (7). The Kronecker limit formula then leads to representations for certain products of Dedekind eta-functions in terms of fundamental units. By (8), these representations allow us to calculate G_n. Our methods extend those of K. G. Ramanathan **[13]** who calculated some of Ramanujan's class invariants but required that $\mathbb{Q}(\sqrt{-n})$ contains only one class per genus. Zhang **[22]**, **[23]** has further extended the method to give rigorous proofs of the invariants of Ramanujan that Watson **[19]** had "empirically" calculated.

Our second method takes Watson's ideas and employs class field theory to put the "empirical" process on a firm foundation **[7]**. It has been further extended by Chan to determine several new invariants **[9]**.

It is highly doubtful that Ramanujan had any acquaintance with Kronecker's limit formula, the arithmetic of quadratic fields, or class field theory. Thus, Ramanujan's ideas still remain hidden behind an opaque curtain.

In this paper, we have made many claims without proofs (as did Ramanujan), but complete proofs or references for all our assertions can be found in **[2]**.

REFERENCES

1. B. C. Berndt, *Ramanujan's Notebooks, Part III*, Springer-Verlag, New York, 1991.
2. B. C. Berndt, *Ramanujan's Notebooks, Part V*, Springer-Verlag, New York, 1997.
3. B. C. Berndt and H. H. Chan, Ramanujan's explicit values for the classical theta-function, *Mathematika* **42** (1995), 278–294.
4. B. C. Berndt and H. H. Chan, Some values for the Rogers-Ramanujan continued fraction *Canad. J. Math.* **47** (1995), 897–914.
5. B. C. Berndt, H. H. Chan, and L.-C. Zhang, Ramanujan's class invariants and cubic continued fraction, *Acta Arith.* **73** (1995), 67–85.
6. B. C. Berndt, H. H. Chan, and L.-C. Zhang, Explicit evaluations of the Rogers-Ramanujan continued fraction, *J. Reine Angew. Math.* **480** (1996), 141–159.
7. B. C. Berndt, H. H. Chan, and L.-C. Zhang, Ramanujan's class invariants, Kronecker's limit formula, and modular equations, *Trans. Amer. Math. Soc.* **349** (1997), 2125–2173.
8. B. C. Berndt and R. A. Rankin, *Ramanujan: Letters and Commentary*, American Mathematical Society, Providence, 1995; London Mathematical Society, London, 1995.
9. H. H. Chan, Ramanujan--Weber class invariant G_n and Watson's empirical process, *J. London Math. Soc.* (to appear).
10. G. Chrystal, *Algebra, 6th ed.*, Chelsea, New York, 1959.
11. D. A. Cox, *Primes of the Form* $x^2 + ny^2$, Wiley, New York, 1989.
12. H. S. Hall and S. R. Knight, *Higher Algebra*, Macmillan, London, 1957.
13. K. G. Ramanathan, Some applications of Kronecker's limit formula, *J. Indian Math. Soc.* **52** (1987), 71–89.
14. S. Ramanujan, Modular equations and approximations to π, *Quart. J. Math. (Oxford)* **45** (1914), 350–372.
15. S. Ramanujan, *Notebooks* (*2 volumes*), Tata Institute of Fundamental Research, Bombay, 1957.
16. S. Ramanujan, *Collected Papers*, Chelsea, New York, 1962.
17. S. Ramanujan, *The Lost Notebook and Other Unpublished Papers*, Narosa, New Delhi, 1988.
18. G. N. Watson, Theorems stated by Ramanujan (XIV): a singular modulus, *J. London Math. Soc.* **6** (1931), 126–132.
19. G. N. Watson, Some singular moduli (I), *Quart. J. Math.* **3** (1932), 81–98.
20. G. N. Watson, Some singular moduli (II), *Quart. J. Math.* **3** (1932) 189–212.
21. H. Weber, *Lehrbuch der Algebra, dritter Band*, Chelsea, New York, 1961.
22. L.-C. Zhang, *Kronecker's limit formula, class invariants and modular equations (II)*, Analytic Number Theory: Proceedings of a Conference in Honor of Heini Halberstam, vol. 2 (B. C. Berndt, H. G. Diamond, and A. J. Hildebrand, eds.), Birkhäuser, Boston, 817–838.
23. L.-C. Zhang, Kronecker's limit formula, class invariants and modular equations (III), *Acta Arith.* (to appear).

Department of Mathematics
University of Illinois
1409 West Green Street
Urbana, Illinois 61801, USA
berndt@math.uiuc.edu

Department of Mathematics
National University of Singapore
Kent Ridge, Singapore 119260
Republic of Singapore
chanhh@math.nus.sg

Department of Mathematics
Southwest Missouri State University
Springfield, MO 65804, USA
liz917f@cnas.smsu.edu

Ramanujan, Taxicabs, Birthdates, ZIP Codes, and Twists

Ken Ono

Dedicated to the memory of S. Ramanujan on the 110th anniversary of his birth.

It is well known that G. H. Hardy travelled in a taxicab numbered 1729 to an English nursing home to visit his bedridden colleague S. Ramanujan. Hardy was disappointed that his cab had such a mundane number, but to his surprise when he mentioned this to Ramanujan, the brilliant Indian mathematician found 1729 to be quite interesting, for it is the smallest integer that has two distinct representations as a sum of two cubes:

$$1729 = 1^3 + 12^3 = 9^3 + 10^3.$$

J. H. Silverman used this famous anecdote to motivate the study of elliptic curves in a recent article [**8**].

Recently I learned that other permutations of the digits 1, 2, 7, and 9 are significant to the Ramanujan story. Two permutations involve Bruce Berndt, the diligent editor of Ramanujan's notebooks. Bruce has devoted most of his professional career to undertaking the daunting task of proving many of Ramanujan's identities (written in notebooks without proofs), but to my surprise his fascination with Ramanujan has profoundly impacted his life outside mathematics. Sonja, Bruce's youngest daughter, was born in 1972. Is this a coincidence, or could it be an example of "Ramanujan family planning"? With more sleuthing I discovered that Bruce's home is in Urbana, Illinois 61802-**7219**. Could there be any truth to the rumor that Bruce paid the postmaster a mere \$12.79 for this vanity zipcode?

In a more serious direction, consider the number 2719, which came to my attention in joint work with K. Soundararajan [**5**]. We begin with the following footnote from Ramanujan's 1916 paper on quadratic forms [**6**, p. 14]:

"... *the even numbers which are not of the form* $x^2 + y^2 + 10z^2$ *are the numbers*

$$4^{\lambda}(16\mu + 6),$$

while the odd numbers that are not of that form, *viz*.,

$$3, 7, 21, 31, 33, 43, 67, 79, 87, 133, 217, 219, 223, 253, 307, 391 \ldots$$

do not seem to obey any simple law.

In view of the list of exceptions, could there be a "simple law" that eluded Ramanujan? After extensive computation, amongst the odd integers two further exceptions emerged, the numbers 679 and of course 2719. A few years ago W. Duke and R. Schulze-Pillot [**3**] (see [**2**] for a survey) made a great breakthrough in the theory of ternary quadratic forms, and from their work it follows that there are only finitely many positive odd integers that are not of the form $x^2 + y^2 + 10z^2$. Could it be that 2719 is the largest such integer?

Unfortunately we do not yet know enough to decide whether or not it is since they obtained no bound beyond which every odd integer is so represented. Although obtaining such a bound appears to be beyond the current state of knowledge, assuming certain Riemann hypotheses, the author and Soundararajan [5] have shown that the only positive odd integers that are not of the form $x^2 + y^2 + 10z^2$ are indeed 679, 2719, and the 16 numbers on Ramanujan's list. Therefore we have very good reason to believe that 2719 is the largest odd integer that is not of the form $x^2 + y^2 + 10z^2$.

Here we explore the special properties that these eighteen integers share. Obviously they are odd numbers n for which there are no integers x, y, and z with $n = x^2 + y^2 + 10z^2$, and we even know that they are all square-free (see [1], [5,Th. 1]) and coprime to 10, but these numbers are linked for much deeper reasons involving some of the most fundamental objects in algebraic number theory and arithmetic geometry. Let me explain.

Following C. F. Gauss, any collection of equivalence classes of ternary quadratic forms that represent the same residue classes (mod M) for every M is called a "genus." In our case, the genus containing Ramanujan's ternary quadratic form $x^2 + y^2 + 10z^2$ contains only one other class, and a representative for this class is the form $2x^2 + 2y^2 + 3z^2 - 2xz$. For convenience define $r_1(n)$ and $r_2(n)$ by

$$r_1(n) := \#\{(x, y, z) \mid x, y, z \in \mathbb{Z},\ x^2 + y^2 + 10z^2 = n\},$$

$$r_2(n) := \#\{(x, y, z) \mid x, y, z \in \mathbb{Z},\ 2x^2 + 2y^2 + 3z^2 - 2xz = n\}.$$

Therefore, Ramanujan wanted a rule for determining those odd n for which $r_1(n) = 0$.

To see the utility in considering both forms together recall Gauss' Three Squares Theorem. Let $h(D)$ denote the number of classes of primitive binary quadratic forms with discriminant D, the usual "class number," and let $r(n)$ denote the number of representations of n by $x^2 + y^2 + z^2$. If $n > 3$ is square-free, then

$$r(n) = \begin{cases} 12h(-4n) & \text{if } n \equiv 1, 2, 5, 6 \pmod 8, \\ 24h(-n) & \text{if } n \equiv 3 \pmod 8. \end{cases}$$

More generally, Gauss obtained formulas for the number of representations of integers by genera, and in the case of Ramanujan's form, if n is a positive square-free integer coprime to 10, then

$$r_1(n)/2 + r_2(n) = h(-40n).$$

Therefore if n is a positive odd integer that is not of the form $x^2 + y^2 + 10z^2$, then

$$r_2(n) = h(-40n). \tag{1}$$

It is also useful to consider the differences $r_1(n) - r_2(n)$. To do so, define

$$f(z) := \frac{1}{4} \sum_{n=1}^{\infty} (r_1(n) - r_2(n)) q^n$$
$$= q - q^3 - q^7 - q^9 + 2q^{13} + \cdots \ (q := e^{2\pi i z} \text{ with } Im(z) > 0). \tag{2}$$

This function f is a "weight 3/2 modular form." An analytic function $m(z)$ on the upper half of the complex plane is a modular form of weight k if for each suitable

matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z})$ there exist roots of unity $\epsilon(d)$ for which

$$m\left(\frac{az+b}{cz+d}\right) = \epsilon(d)(cz+d)^k m(z).$$

To study $r_1(n) - r_2(n)$ we employ the *Shimura lift* [7], a beautiful correspondence between certain half-integral weight modular forms and integral weight modular forms. In this case if integers $A(n)$ are defined by

$$\sum_{n=1}^{\infty} \frac{A(n)}{n^s} := \frac{1}{4} \cdot \left(\sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} \right) \cdot \left(\sum_{m=1}^{\infty} \frac{r_1(m^2) - r_2(m^2)}{m^s} \right)$$

where χ denotes the Legendre-Kronecker quadratic character $\chi(n) := \left(\frac{-10}{n}\right)$, and if q and z are as in (2), then

$$\begin{aligned} F(z) &:= \sum_{n=1}^{\infty} A(n) q^n = q \prod_{n=1}^{\infty} (1-q^{2n})^2 (1-q^{10n})^2 \\ &= q - 2q^3 - q^5 + 2q^7 + q^9 + 2q^{13} + \cdots \end{aligned} \tag{3}$$

is a weight 2 modular form.

The modular form $F(z)$ provides an example of the celebrated Shimura-Taniyama Conjecture, whose proof in special cases by A. Wiles yields Fermat's Last Theorem. The conjecture asserts that the coefficients of certain weight 2 modular forms, the $A(n)$, equal the coefficients of L-functions of elliptic curves. In this case let E denote the elliptic curve over the rational numbers

$$E: \; y^2 = x^3 + x^2 + 4x + 4.$$

For each odd prime p let $N(p)$ denote the number of pairs, $x \pmod p$, $y \pmod p$, that satisfy the congruence

$$y^2 \equiv x^3 + x^2 + 4x + 4 \pmod p.$$

If $c(p) := p - N(p)$, then the Hasse-Weil L-function $L(E,s)$ is defined by the following product over all the odd primes:

$$\begin{aligned} L(E,s) &:= \sum_{n=1}^{\infty} \frac{c(n)}{n^s} = \prod_{p \neq 2} \frac{1}{1 - c(p)p^{-s} + p^{1-2s}} \\ &= 1 - \frac{2}{3^s} - \frac{1}{5^s} + \frac{2}{7^s} + \frac{1}{9^s} + \frac{2}{13^s} + \cdots. \end{aligned} \tag{4}$$

By comparing (3) and (4) one sees that $A(n) = c(n)$ for each $n \leq 13$. In fact this equality holds for every positive integer n, and is an example of the phenomenon described by the Shimura-Taniyama Conjecture.

Are these observations relevant to Ramanujan's query? They are, and the answer lies in the work of J.-L. Waldspurger **[10]** who provided a very deep and beautiful interpretation of Shimura's lift. In our case let n be a positive odd square-free integer, and define the $-10n$ quadratic twist $L(E(-10n),s)$ by

$$L(E(-10n),s) := \prod_{p \neq 2,5} \frac{1}{1 - c(p)\left(\frac{-10n}{p}\right)p^{-s} + p^{1-2s}}.$$

This is the L-function for the elliptic curve $E(-10n)$

$$E(-10n): \; y^2 = x^3 - 10nx^2 + 400n^2x - 4000n^3.$$

If

$$\Omega := \int_{10}^{\infty} \frac{1}{\sqrt{x^3 - 10x^2 + 400x - 4000}}\, dx \sim 0.7195\ldots,$$

then for every odd square-free integer $n \neq 5$ Waldspurger's theorem implies

$$(r_1(n) - r_2(n))^2 = \frac{4\sqrt{n}}{\Omega} \cdot L(E(-10n), 1).$$

Therefore, by (1), if n is a positive odd integer that is not of the form $x^2 + y^2 + 10z^2$, then

$$h^2(-40n) = \frac{4\sqrt{n}}{\Omega} \cdot L(E(-10n), 1). \tag{5}$$

Although (5) is a "law" that the odd integers not of the form $x^2 + y^2 + 10z^2$ obey, it certainly is not a simple one. However, its formulation is particularly intriguing.

First we recall some facts about elliptic curves. Let C denote the set of rational points (x, y) satisfying

$$C:\ y^2 = x^3 + ax^2 + bx + c$$

where a, b and c are fixed rational numbers. If the discriminant of $x^3 + ax^2 + bx + c$ is non-zero, then Mordell proved that C, including a "point at infinity," forms a finitely generated abelian group whose group law is a "chord-tangent" law (see **[9]**). Therefore,

$$C \cong C_{\text{torsion}} \times \mathbb{Z}^r$$

where C_{torsion}, the torsion subgroup of C, is a finite abelian group, and the rank r is a non-negative integer. Note that C has finitely many points if and only if $r = 0$. Quite a bit is known about C_{torsion}. By a theorem of Mazur it is known that C_{torsion} satisfies

$$C_{\text{torsion}} \in \begin{cases} \mathbb{Z}m & |\text{where } 1 \le m \le 10, \text{ or } m = 12, \\ \mathbb{Z}2 \times \mathbb{Z}2m & |\text{where } 1 \le m \le 4 \end{cases}$$

($\mathbb{Z}d$ denotes the cyclic group with d elements), and with this classification it is fairly easy to deduce C_{torsion} for any given C.

Computing r is a more difficult question, and although one can typically compute r in practice, the problem in general remains open. In part these problems revolve around the Birch and Swinnerton-Dyer Conjecture, which asserts that the analytic behavior of $L(C, s)$ at $s = 1$ predicts the structure of C, in particular r. In its weakest form the conjecture asserts that $L(C, s)$ has an analytic continuation to the entire complex plane, and that r equals the order of vanishing at $s = 1$ of $L(C, s)$. In particular, C has finitely many points precisely when $L(C, 1) \neq 0$.

For a "modular" elliptic curve C, one satisfying the Shimura-Taniyama Conjecture, V. Kolyvagin **[4]** proved that C has finitely many points if $L(C, 1) \neq 0$. Therefore by the positivity of $h(-40n)$, (5), and Kolyvagin's theorem, if n is a positive odd integer that is not of the form $x^2 + y^2 + 10z^2$, then $E(-10n)$ has finitely many points. In fact if n equals 679, 2719, or any of the 16 numbers on Ramanujan's list, then the only rational point (x, y) on $E(-10n)$

$$y^2 = x^3 - 10nx^2 + 400n^2x - 4000n^3,$$

is $(10n, 0)$.

In its full strength the Birch and Swinnerton-Dyer Conjecture predicts even more. If $L(C,1) \neq 0$ the conjecture predicts that $L(C,1)$ is an explicit real multiple of the order of $Ш(C)$, the Tate-Shafarevich group of C, which measures the extent to which the "local-global principle" fails for an elliptic curve C. Recall that a conic has a point with rational coordinates precisely when it contains a point with real coordinates and a point with coordinates that are p-adic numbers for every prime p. However this is not true for elliptic curves. In a famous example, E. Selmer noted that there are no non-trivial rational points on

$$3x^3 + 4y^3 + 5z^3 = 0,$$

even though it has points over every field of p-adic numbers. The Tate-Shafarevich group measures the failure of this principle.

In our case if n is 679, 2719, or one of the 16 integers on Ramanujan's list, then (5) and the Birch and Swinnerton-Dyer Conjecture imply

$$h^2(-40n) = 4^{t(n)+1}|Ш(E(-10n))| \tag{6}$$

where $t(n)$ denotes the number of prime factors of n.

Just as Tate-Shafarevich groups measure the obstruction to the "local-global" principle for elliptic curves, the set of classes of discriminant D primitive binary quadratic forms, denoted by $CL(D)$, measures an obstruction. The set $CL(D)$ is an abelian group with order $h(D)$ that is isomorphic to the "ideal class group" of $\mathbb{Q}(\sqrt{D})$, and it measures the extent to which unique factorization fails in the ring of integers of $\mathbb{Q}(\sqrt{D})$. For instance, in the ring of integers of $\mathbb{Q}(\sqrt{-40})$, numbers of the form $a + b\sqrt{-10}$ with $a, b \in \mathbb{Z}$, the integer 14 does not factor uniquely into irreducibles since it has the following factorizations:

$$14 = 2 \cdot 7 = (2 + \sqrt{-10}) \cdot (2 - \sqrt{-10}).$$

By Gauss' genus theory the index of the subgroup

$$CL^2(-40n) := \{\alpha^2 | \alpha \in CL(-40n)\}$$

in $CL(-40n)$ is $2^{t(n)+1}$. Therefore for the known odd integers not of the form $x^2 + y^2 + 10z^2$, (6) and the Birch and Swinnerton-Dyer Conjecture imply the following tantalizing equality relating class groups and Tate-Shafarevich groups:

$$|CL^2(-40n) \times CL^2(-40n)| = |Ш(E(-10n))|. \tag{7}$$

From our discussion, Ramanujan's search for a "simple law" leads to several deep theorems and conjectures in arithmetic geometry. To reiterate, if n equals 679, 2719, or one of the integers on Ramanujan's list, then using Shimura's lift, the Shimura-Taniyama correspondence, and the works of Kolyvagin and Waldspurger, we have obtained the following gems:

(i) There are no rational numbers x and y with $y \neq 0$ for which

$$y^2 = x^3 - 10nx^2 + 400n^2x - 4000n^3.$$

(ii) Assuming the Birch and Swinnerton-Dyer Conjecture,

$$|CL^2(-40n) \times CL^2(-40n)| = |Ш(E(-10n))|.$$

There are a few other ternary forms that also have such elegant properties, but most do not. This illustrates again how Ramanujan's deep insight continues to thrive beyond his centenary. By the way, Ramanujan's lifespan was 1887-**1920**.

ACKNOWLEDGMENTS. The author thanks B. C. Berndt and J. H. Silverman for their helpful comments during the preparation of this article, and he also thanks the Institute for Advanced Study for their hospitality during the 1995–1996 and 1996–1997 academic years. The author's research is supported in part by NSF grants DMS-9304580 and DMS-9508976.

REFERENCES

1. J. Benham and J. Hsia, Spinor equivalence of quadratic forms *J. Number Th.* **17** (1983), 337–342.
2. W. Duke, Some old problems and new results about quadratic forms *Notices Amer. Math. Soc.* **44** (1997), 190–196.
3. W. Duke and R. Schulze-Pillot, Representations of integers by positive ternary quadratic forms and equidistribution of lattice points on ellipsoids, *Invent. Math.* **99** (1990) 49–57.
4. V.A. Kolyvagin, Finiteness of $E(\mathbb{Q})$ and the Tate-Shafarevich group of $E(\mathbb{Q})$ for a subclass of Weil curves (Russian), *Izv. Akad. Nauk, USSR, ser. Matem.* **52** (1988), 522–540.
5. K. Ono and K. Soundararajan, Ramanujan's ternary quadratic form, *Invent. Math.* **130** (1997) 415–454.
6. S. Ramanujan, On the expression of a number in the form $ax^2 + by^2 + cz^2 + du^2$, *Proc. Cambridge Phil. Soc.* **19** (1916), 11–21.
7. G. Shimura, On modular forms of half-integral weight, *Ann. Math.* **97** (1973), 440–481.
8. J. H. Silverman, Taxicabs and sums of two cubes, *Amer. Math. Monthly* **100** (1993), 331–40.
9. J. H. Silverman and J. Tate, *Rational points on elliptic curves*, Springer Verlag, New York, New York, 1992.
10. J.-L. Waldspurger, Sur les coefficients de Fourier des formes modulaires de poids demi-entier, *J. Math. Pures et Appl.* **60** (1981), 375–484.

Department of Mathematics
Penn State University
University Park, Pennsylvania 16802
ono@math.psu.edu

Whenever I am angry or depressed, I pull down [his] collected papers from the shelf and take a quiet stroll in Ramanujan's garden. I recommend this therapy to all of you who suffer from headaches or jangled nerves. And Ramanujan's papers are not only a good therapy for headaches. They also are full of beautiful ideas which may help you to do more interesting mathematics.

Freeman Dyson, *Selected Papers*,
American Mathematical Society, Providence, 1996, p. 205
Contributed by Timothy Chambers

Simplicity and Surprise in Ramanujan's "Lost" Notebook

George E. Andrews

1. INTRODUCTION. In 1979, I wrote an introduction [2] to Ramanujan's "Lost" Notebook [7]. In that introduction I provided a short history of my connection with this amazing document as well as a sampling of some of the results.

Suffice it to say that the "Lost" Notebook contains a substantial collection of the discoveries the Indian genius, Ramanujan, made during 1919–1920, the last year of his short life. Furthermore, nothing was published on any of the formulas in the "Lost" Notebook until the appearance of [2] in 1979.

One of the formulas presented in the introduction was [7; p. 47], [2; p. 90, eq. (1.3)]

$$\left(\cfrac{1}{1+\cfrac{q}{1+\cfrac{q^2}{1+\cfrac{q^3}{1+\cfrac{q^4}{\ddots}}}}}\right)^3 = \frac{\sum_{n=0}^{\infty} q^{5n^2+4n}\dfrac{1+q^{5n+2}}{1-q^{5n+2}} - \sum_{n=0}^{\infty} q^{5n^2+6n+1}\dfrac{1+q^{5n+3}}{1-q^{5n+3}}}{\sum_{n=0}^{\infty} q^{5n^2+2n}\dfrac{1+q^{5n+1}}{1-q^{5n+1}} - \sum_{n=0}^{\infty} q^{5n^2+8n+3}\dfrac{1+q^{5n+4}}{1-q^{5n+4}}} \tag{1.1}$$

While cubing a continued fraction may be surprising, it is certainly not simple.

There are, however, much less daunting results in the "Lost" Notebook. Here are two that illustrate simplicity and surprise [7; p. 31]:

$$\begin{aligned} &1 + \frac{q}{(1-q)(1-q^2)} + \frac{q^2}{(1-q)(1-q^2)(1-q^3)(1-q^4)} \\ &\quad + \frac{q^3}{(1-q)(1-q^2)(1-q^3)(1-q^4)(1-q^5)(1-q^6)} + \cdots \\ &\quad = \frac{(1+q^{11}+q^{13}+\cdots) - q(1+q^5+q^{19}+\cdots)}{1-2q+2q^4-2q^9+2q^{16}-2q^{25}+\cdots} \end{aligned} \tag{1.2}$$

and

$$\begin{aligned} &1 + \frac{q}{(1+q)(1+q^2)} + \frac{q^2}{(1+q)(1+q^2)(1+q^3)(1+q^4)} \\ &\quad + \frac{q^3}{(1+q)(1+q^2)(1+q^3)(1+q^4)(1+q^5)(1+q^6)} + \cdots \\ &\quad = (1-q^{11}+q^{13}-\cdots) + q(1-q^5+q^{19}-\cdots). \end{aligned} \tag{1.3}$$

The sequence $0, 11, 13, \ldots$ is $\{12n^2 + n\}_{n=-\infty}^{\infty}$ and $0, 5, 19, \ldots$ is $\{12n^2 + 7n\}_{n=-\infty}^{\infty}$. Of course $0, 1, 4, 9, 16, \ldots$ is $\{n^2\}_{n=0}^{\infty}$.

Please note the unreasonable fact that the passage from (1.2) to (1.3) consists of changing various signs and dropping the denominator on the right-hand side of (1.2). The implausibility of this transformation is a standard Ramanujan surprise.

One may, of course, object that (1.2) and (1.3) are not all that simple. I will try to convince you otherwise in Section 2 with an account of related formulas that Euler found (or could have found) and that can be proved easily by mathematical induction. In Sections 3 and 4 we show why Ramanujan's discoveries turn out to be surprisingly deeper than those of Euler's, and while the formulas are simply stated their proofs are not that easy, another surprise. In Section 5, I briefly describe the number-theoretic implication of these results, and I conclude with a sketch of related discoveries by Ramanujan.

2. BACKGROUND. We consider two formulas in this section. They look like slightly simpler cousins of (1.2) and (1.3). They are, it turns out, easily proved.

$$\begin{aligned} &1 + \frac{q}{1-q} + \frac{q^2}{(1-q)(1-q^2)} + \frac{q^3}{(1-q)(1-q^2)(1-q^3)} + \cdots \\ &\quad = \frac{1}{(1-q)(1-q^2)(1-q^3)(1-q^4)(1-q^5)\cdots} \end{aligned} \tag{2.1}$$

$$\begin{aligned} &1 + \frac{q}{1+q} + \frac{q^2}{(1+q)(1+q^2)} + \frac{q^3}{(1+q)(1+q^2)(1+q^3)} + \cdots \\ &\quad = 2 - (1-q)(1-q^3)(1-q^5)(1-q^7)\cdots \end{aligned} \tag{2.2}$$

Identity (2.1) is a special case of a formula due to Euler [**1**; p. 19, eq. (2.2.5)], and (2.2) can be deduced from a special case of Heine's transformation of q-hypergeometric series [**1**; p. 19, Cor. 2.3]. However, each is a limiting case of two formulas whose proof is an immediate mathematical induction exercise. Namely

$$1 + \sum_{j=1}^{n} \frac{q^j}{(1-q)(1-q^2)\cdots(1-q^j)} = \frac{1}{(1-q)(1-q^2)\cdots(1-q^n)}, \tag{2.3}$$

$$1 + \sum_{j=1}^{n} \frac{q^j}{(1+q)(1+q^2)\cdots(1+q^j)} = 2 - \frac{1}{(1+q)(1+q^2)\cdots(1+q^n)}. \tag{2.4}$$

The mathematical induction proof of (2.3) hinges on the identity

$$\begin{aligned} &\frac{1}{(1-q)(1-q^2)\cdots(1-q^n)} - \frac{1}{(1-q)(1-q^2)\cdots(1-q^{n-1})} \\ &\quad = \frac{1-(1-q^n)}{(1-q)(1-q^2)\cdots(1-q^n)} = \frac{q^n}{(1-q)(1-q^2)\cdots(1-q^n)}, \end{aligned} \tag{2.5}$$

while the proof of (2.4) relies on

$$\begin{aligned} &\left(2 - \frac{1}{(1+q)(1+q^2)\cdots(1+q^n)}\right) - \left(2 - \frac{1}{(1+q)(1+q^2)\cdots(1+q^{n-1})}\right) \\ &\quad = \frac{(1+q^n)-1}{(1+q)(1+q^2)\cdots(1+q^n)} = \frac{q^n}{(1+q)(1+q^2)\cdots(1+q^n)}. \end{aligned} \tag{2.6}$$

A complete detailed proof would be a good exercise for a college algebra class.

If we let n tend to infinity in (2.3), we deduce (1.2). If we let n tend to infinity in (2.4), we find

$$1+\frac{q}{1+q}+\frac{q^2}{(1+q)(1+q^2)}+\frac{q^3}{(1+q)(1+q^2)(1+q^3)}+\cdots$$
$$=2-\frac{1}{(1+q)(1+q^2)(1+q^3)(1+q^4)\cdots}, \tag{2.7}$$

and (2.7) is equivalent to (2.2) because of Euler's famous infinite product identity [**1**; p. 5, eq. (1.2.5)]:

$$\frac{1}{(1+q)(1+q^2)(1+q^3)(1+q^4)(1+q^5)\cdots} \tag{2.8}$$
$$=\frac{(1-q)}{(1-q^2)}\cdot\frac{(1-q^2)}{(1-q^4)}\cdot\frac{(1-q^3)}{(1-q^6)}\cdot\frac{(1-q^4)}{(1-q^8)}\cdot\frac{(1-q^5)}{(1-q^{10})}\cdots$$
$$=(1-q)(1-q^3)(1-q^5)\cdots$$

What could be simpler?

By means of two famous results, the first by Euler [**1**; p. 11, eq. (1.3.1)] and the second by Gauss [**1**; p. 23, eq. (2.2.12)],

$$\prod_{n=1}^{\infty}(1-q^n)=1+\sum_{n=1}^{\infty}(-1)^n q^{n(3n-1)/2}(1+q^n), \tag{2.9}$$

and

$$\prod_{n=1}^{\infty}\frac{(1-q^n)}{(1+q^n)}=1+2\sum_{n=1}^{\infty}(-1)^n q^{n^2}, \tag{2.10}$$

we can recast (2.1) and (2.2) so that they look much more like (1.2) and (1.3). Namely by (2.9)

$$1+\frac{q}{1-q}+\frac{q^2}{(1-q)(1-q^2)}+\frac{q^3}{(1-q)(1-q^2)(1-q^3)}+\cdots$$
$$=\frac{1}{1-q-q^2+q^5+q^7-q^{12}-q^{15}+\cdots}, \tag{2.11}$$

and by (2.9) and (2.10)

$$1+\frac{q}{1+q}+\frac{q^2}{(1+q)(1+q^2)}+\frac{q^3}{(1+q)(1+q^2)(1+q^3)}+\cdots$$
$$=2-\frac{1-2q+2q^4-2q^9+2q^{16}-2q^{25}+\cdots}{1-q-q^2+q^5+q^7-q^{12}-q^{15}+\cdots} \tag{2.12}$$

3. RAMANUJAN'S FORMULA (1.2). I direct your attention to the fact that Ramanujan's series on the left sides of (1.2) and (1.3) are obtained from the corresponding series in (2.1) and (2.2) by doubling the number of factors in the denominator, surely no big deal.

However, it is a big deal! For starters, we no longer have simple representations like (2.3) and (2.4) for the partial sums. For example,

$$1 + \frac{q}{(1-q)(1-q^2)} + \frac{q^2}{(1-q)(1-q^2)(1-q^3)(1-q^4)}$$
$$= \frac{1 - q^4 + q^5 - q^9 - q^{10}}{1 - q - q^2 + 2q^5 - q^8 - q^9 + q^{10}}, \tag{3.1}$$

a mess that gives no hint of the right-hand side of (1.3). Indeed, failure has met every attempt I know of to prove (1.2) and (1.3) simply. The only proofs I know rely on results deeper than (2.1) and (2.2). To prove (1.2), we require Euler's full identity [**1**; p. 19, eq. (2.2.5)]

$$1 + \sum_{j=1}^{\infty} \frac{z^j}{(1-q)(1-q^2)\cdots(1-q^j)} = \frac{1}{(1-z)(1-zq)(1-zq^2)\cdots}, \tag{3.2}$$

Jacobi's Triple Product Identity [**1**; p. 21. eq. (2.2.10)], [**6**; p. 12, eq. (1.6.1)] one of the cornerstones of elliptic theta function theory

$$\sum_{n=-\infty}^{\infty} z^n q^{n(n-1)/2} = \prod_{m=0}^{\infty} (1 - q^{m+1})(1 + zq^m)(1 + z^{-1}q^{m+1}), \tag{3.3}$$

and the lesser known but elegant Quintuple Product Identity [**6**; p. 134]

$$\sum_{n=-\infty}^{\infty} z^{3n} q^{n(3n-1)/2}(1 + zq^n)$$
$$= \prod_{m=0}^{\infty} (1 - q^{m+1})(1 + zq^m)(1 + z^{-1}q^{m+1}/z)(1 - z^2q^{2m+1})(1 - q^{2m+1}/z^2). \tag{3.4}$$

Hence

$$1 + \sum_{n=1}^{\infty} \frac{q^n}{(1-q)(1-q^2)\cdots(1-q^{2n})}$$
$$= 1 + \frac{1}{2}\sum_{n=1}^{\infty} \frac{q^{n/2}(1 + (-1)^n)}{(1-q)(1-q^2)\cdots(1-q^n)} \text{ (the terms with odd } n \text{ are zero)}$$
$$= \frac{1}{2(1-q^{1/2})(1-q^{3/2})(1-q^{5/2})\cdots} + \frac{1}{2(1+q^{1/2})(1+q^{3/2})(1+q^{5/2})\cdots}$$

(by two applications of (3.2), the first with $z = q^{1/2}$, the second with $z = -q^{1/2}$)

$$= \frac{1}{2}\left\{\frac{(1+q^{1/2})(1+q^{3/2})(1+q^{5/2})\cdots + (1-q^{1/2})(1-q^{3/2})(1-q^{5/2})\cdots}{(1-q)(1-q^3)(1-q^5)\cdots}\right\}$$

(by the same algebraic cancellation used in (2.8))

$$= \frac{1}{2}\left\{\frac{\displaystyle\prod_{m=0}^{\infty}(1-q^{2m+2})(1+q^{2m+1/2})(1+q^{2m+3/2}) + \prod_{m=0}^{\infty}(1-q^{2m+2})(1-q^{2m+1/2})(1-q^{2m+3/2})}{\displaystyle\left(\prod_{m=0}^{\infty}(1-q^{2m+2})\right)(1-q)(1-q^3)(1-q^5)\cdots}\right\}$$

$$= \frac{1}{2}\left\{\frac{\displaystyle\sum_{n=-\infty}^{\infty} q^{n^2+n/2} + \sum_{n=-\infty}^{\infty}(-1)^n q^{n^2+n/2}}{\displaystyle\prod_{m=0}^{\infty}(1-q^{2m+2})\prod_{m=0}^{\infty}(1-q^{2m+1})}\right\}$$

$$= \frac{\displaystyle\sum_{n=-\infty}^{\infty} q^{4n^2+n}}{\displaystyle\prod_{m=0}^{\infty}(1-q^{2m+2})\prod_{m=0}^{\infty}(1-q^{2m+1})} \quad \text{(the terms with } n \text{ odd cancelled)}$$

$$= \frac{\displaystyle\prod_{n=0}^{\infty}(1-q^{8n+8})(1+q^{8n+3})(1+q^{8n+5})}{\displaystyle\prod_{m=0}^{\infty}(1-q^{2m+2})\prod_{m=0}^{\infty}(1-q^{2m+1})}$$

$$= \frac{\displaystyle\prod_{n=0}^{\infty}(1-q^{8n+8})(1-q^{8n+1})(1-q^{8n+7})(1-q^{16n+6})(1-q^{16n+10})}{\displaystyle\prod_{n=0}^{\infty}(1-q^{2n+2})(1-q^{2n+1})^2}$$

(here we multiplied numerator and denominator by $\prod_{m=0}^{\infty}(1-q^{2m+1})$ and then rewrote the numerator using $\prod_{m=0}^{\infty}(1-q^{2m+1}) = \prod_{m=0}^{\infty}(1-q^{8m+1})(1-q^{8m+3})(1-q^{8m+5})(1-q^{8m+7})$)

$$= \frac{\displaystyle\sum_{n=-\infty}^{\infty}\left(q^{12n^2-n} - q^{12n^2+7n+1}\right)}{1-2q+2q^4-2q^9+2q^{16}-2q^{25}+\cdots}.$$

The last expression is finally the right-hand side of (1.2). The final numerator is obtained from the penultimate numerator by invoking (3.4) with q replaced by q^8 and then z replaced by $-q$. The final denominator is obtained from the penultimate denominator by invoking (3.3) with q replaced by q^2 and z then replaced by -1.

While no one rediscovered (1.2) prior to the unearthing of the "Lost" Notebook, Leonard Carlitz [**4**; eq. (24)] did prove a formula equivalent to (1.2). His identity is essentially the above stopping at the antepenultimate line.

As hard as it may be to believe, identity (1.3) is an even tougher nut than (1.2), as we shall see in the next section.

4. RAMANUJAN'S FORMULA (1.3). Our starting point for (1.3) is a formula also taken from the "Lost" Notebook [**7**; p. 37] ([**3**; p. 137, eq. (1.1)], cf. [**5**; Ch. 1, §7]). This formula is substantially more difficult to prove than any of the background

formulas used in Section 3.

$$1+\sum_{n=1}^{\infty}\frac{q^n}{(1+aq)(1+a^{-1}q)(1+aq^2)(1+a^{-1}q^2)\cdots(1+aq^n)(1+a^{-1}q^n)}$$

$$=(1+a)\sum_{n=0}^{\infty}a^{3n}q^{n(3n+1)/2}(1-a^2q^{2n+1})-\frac{a\sum_{n=0}^{\infty}(-1)^n a^{2n}q^{n(n+1)/2}}{\prod_{j=1}^{\infty}(1+aq^j)(1+a^{-1}q^j)}. \tag{4.1}$$

Now set $a=i(=\sqrt{-1})$ in (4.1) and take real parts of both sides

$$1+\sum_{n=1}^{\infty}\frac{q^n}{(1+q^2)(1+q^4)\cdots(1+q^{2n})}=\mathrm{Re}\left\{(1+i)\sum_{n=1}^{\infty}i^{3n}q^{n(3n+1)/2}(1+q^{2n+1})\right.$$

$$=\sum_{n=0}^{\infty}(-1)^n q^{n(6n+1)}(1+q^{4n+1})+\sum_{n=0}^{\infty}(-1)^n q^{(2n+1)(3n+2)}(1+q^{4n+3}) \tag{4.2}$$

I need hardly remark that the passage from (4.1) to (4.2) has been a dramatic simplification. Whereas in (4.1) the right-hand side was a combination of infinite series and infinite products, the right-hand side of (4.2) is now a power series in q whose only coefficients are 0 and ± 1.

From (4.2), identity (1.3) follows easily. Let us call the left-hand side of (1.3) $f(q)$ and the left-hand side of (4.2) $h(q)$. Then

$$\begin{aligned}
f(q^2) &= 1+\sum_{n=1}^{\infty}\frac{q^{2n}}{(1+q^2)(1+q^4)\cdots(1+q^{4n})}\\
&= 1+\frac{1}{2}\sum_{n=1}^{\infty}\frac{q^n(1+(-1)^n)}{(1+q^2)(1+q^4)\cdots(1+q^{2n})}\\
&= \frac{1}{2}(h(q)+h(-q))\\
&= \text{Even part of}\left(\sum_{n=0}^{\infty}(-1)^n q^{n(6n+1)}(1+q^{4n+1})\right.\\
&\qquad\left.+\sum_{n=0}^{\infty}(-1)^n q^{(2n+1)(3n+2)}(1+q^{4n+3})\right)\\
&\quad (\text{by } (4.2))\\
&= \sum_{n=0}^{\infty}q^{24n^2+2n}-\sum_{n=0}^{\infty}q^{24n^2+34n+12}\\
&\quad -\sum_{n=0}^{\infty}q^{24n^2+14n+2}+\sum_{n=0}^{\infty}q^{24n^2+46n+22}.
\end{aligned}$$

Hence

$$f(q)=\sum_{n=0}^{\infty}q^{12n^2+n}(1-q^{22n+11})+q\sum_{n=0}^{\infty}q^{12n^2+7n}(1-q^{10n+5}),$$

which is (1.3).

5. APPLICATION TO PARTITIONS. We could easily derive many number-theoretic results from the identities developed in Sections 3 and 4. We limit ourselves to one of the simplest.

In additive number theory, partitions refer to unordered decompositions of integers into sums of positive integers. For example, the five partitions of 4 are $4, 3+1, 2+2, 2+1+1, 1+1+1+1$. Note the adjective "unordered": $1+2+1$ and $2+1+1$ are considered identical.

Let $P_a(n)$ ($a = 1$ or 3) denote the number of partitions of n into parts where only the largest part appears an odd number of times and the total number of parts is congruent to $a \bmod 4$.

For example, there are four partitions of 6 in which only the largest part appears an odd number of times: $6, 4+1+1, 2+2+2, 2+1+1+1+1$. So $P_1(6) = 2 = P_3(6)$.

There are thirteen partitions of 12 in which only the largest part appears an odd number of times: 12, $10+1+1$, $8+2+2$, $8+1+1+1+1$, $6+2+2+1+1$, $6+1+1+1+1+1+1$, $4+4+4$, $4+2+2+2+2$, $4+2+2+1+1+1+1$, $4+1+1+1+1+1+1+1+1$, $2+2+2+2+2+1+1$, $2+2+2+1+1+1+1+1+1$, $2+1+1+1+1+1+1+1+1+1+1$. So $P_1(12) = 6$ and $P_3(12) = 7$.

The following theorem reveals that $P_1(n)$ and $P_3(n)$ never differ by more than 1.

Theorem. *For each $n > 0$,*

$$P_1(n) - P_3(n) = \begin{cases} 0 & \text{if } n \neq j(3j-1)/2 \\ (-1)^j & \text{if } n = \begin{cases} 6j^2 + j, \\ 6j^2 + 5j + 1, \\ 6j^2 + 7j + 2, \\ \text{or} \\ 6j^2 + 11j + 5. \end{cases} \end{cases} \tag{5.1}$$

Proof: The elementary techniques of partition theory [**1**; Ch. 1] reveal that the coefficient of $z^m q^N$ in

$$\mathscr{P}(z,q) = \sum_{n=1}^{\infty} \frac{zq^n}{(1-z^2q^2)(1-z^2q^4)\cdots(1-z^2q^{2n})} \tag{5.2}$$

is the number of partitions of N into m parts wherein only the largest part appears an odd number of times. Since $P(z,q)$ is an odd function of z, we see that

$$\begin{aligned} \sum_{n=1}^{\infty} (P_1(n) - P_3(n))q^n &= \frac{1}{i}\mathscr{P}(i,q) \\ &= \sum_{n=1}^{\infty} \frac{q^n}{(1+q^2)(1+q^4)\cdots(1+q^{2n})}, \end{aligned} \tag{5.3}$$

and the theorem now follows immediately by comparison with (4.2).

6. CONCLUSION. The biggest surprise is how closely the formulas in (1.2) and (1.3) seem to be related. The numerators on the right-hand sides are essentially the same except for a few sign changes, while (1.2) has a classical theta series in the

denominator ($\sum_{-\infty}^{\infty} q^{n^2} = \vartheta_3(0, q)$), and (1.2) has no denominator. Yet these two formulas have very, very different proofs and are progressively harder than (2.1) and (2.2).

There are companions to (1.2) and (1.3) that may be proved in very similar ways. Namely [7; p. 31] (cf. [4, eq. (25)])

$$\sum_{n=0}^{\infty} \frac{q^n}{(1-q)(1-q^2)\cdots(1-q^{2n+1})} = \frac{(1+q^7+q^{17}+\cdots) - q^2(1+q+q^{23}+\cdots)}{1-2q+2q^4-2q^9+2q^{16}-2q^{25}+\cdots} \tag{6.1}$$

and

$$\sum_{n=0}^{\infty} \frac{q^n}{(1+q)(1+q^2)\cdots(1+q^{2n+1})} = (1-q^7+q^{17}-\cdots) + q^2(1-q+q^{23}-\cdots) \tag{6.2}$$

Finally, this paper is just another sample (as was [2]). There has been an extensive series of papers written on the "Lost" Notebook. Many of these are chronicled in [7; pp. xi–xxv]. Bruce Berndt and I are preparing a fully edited account of the work in the "Lost" Notebook, a project that will take some time.

ACKNOWLEDGMENT. Partially supported by National Science Foundation Grant DMS-9501101.

REFERENCES

1. G. E. Andrews, *The Theory of Partitions*, Encyclopedia of Mathematics and Its Applications, Vol. 2, Addison-Wesley, Reading, 1976 [Reissued: Cambridge Univ. Press, Cambridge, 1985].
2. G. E. Andrews, An introduction to Ramanujan's "Lost" notebook, *Amer. Math. Monthly* **86** (1979), 89–109.
3. G. E. Andrews, Ramanujan's "lost" notebook I. Partial θ-functions, *Adv. Math.* **41** (1981), 137–172.
4. L. Carlitz, Note on some continued fractions of the Rogers-Ramanujan type, *Duke Math. J.* **32** (1965), 713–720.
5. N. J. Fine, Basic Hypergeometric Series and Applications, Math. Surveys and Monographs Vol. 27, Amer. Math. Society, Providence, 1988.
6. G. Gasper and M. Rahman, *Basic Hypergeometric Functions*, Encyclopedia of Mathematics and Its Applications, Vol. 35, Cambridge Univ. Press, Cambridge, 1990.
7. S. Ramanujan, *The Lost Notebook and Other Unpublished Papers*, intro. by G. E. Andrews, Narosa Publishing House, New Delhi, 1987.

The Pennsylvania State University
University Park, PA 16802
andrews@math.psu.edu

The Catalan Numbers, the Lebesgue Integral, and 4^{n-2}

Wen-Jin Woan, Lou Shapiro, and D. G. Rogers

In this note we give a new proof of a theorem that is simple to state, rather elegant, and little known. The proof we present is novel in that it involves a use of generating functions that is akin to Lebesgue's parable about a shopkeeper totalling the receipts at the end of the day. The Riemann integral corresponds to just adding up the receipts in order but the Lebesgue integral corresponds to first sorting the receipts by denomination, totalling the amount of each denomination, and then adding up the subtotals.

A *path pair* of length n is a pair of paths that start at the origin, consist of n unit steps and meet again for the first time after n steps. All steps in these paths go East or North. A path pair may also be called a *parallelo-polyomino.* Figure 1 illustrates one such path pair of length 8.

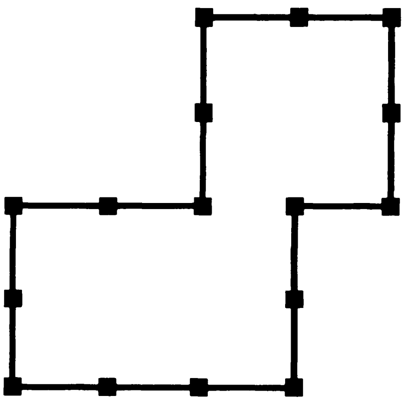


Figure 1

The number of path pairs of length n is C_{n-1}, where $C_0, C_1, C_2, \dots$ are the *Catalan numbers*. The first two proofs of this are by Levine **[9]** and Pólya **[10]**. Our main theorem is that the total area of these C_{n-1} path pairs is 4^{n-2}. After a quick review of the Catalan numbers, we present brief proofs of both facts; the proof of the 4^{n-2} result is new. A far different proof is alluded to in the paper of Fürlinger and Hofbauer [4] and is attributed to a graduate student by the name of Schwarzler. A different approach to a closely related result is given in [**6**]. An unsolved problem is whether the C_{n-1} path pairs of length n can be used as tiles to cover a $2^{n-2} \times 2^{n-2}$ checkerboard. The case $n = 5$ makes an amusing puzzle. The 14 possible shapes are illustrated in Figure 2. To play the game, duplicate and enlarge the 14 pieces by, say, a factor of four, cut them out, and then try to arrange them so that they cover an 8×8 board.

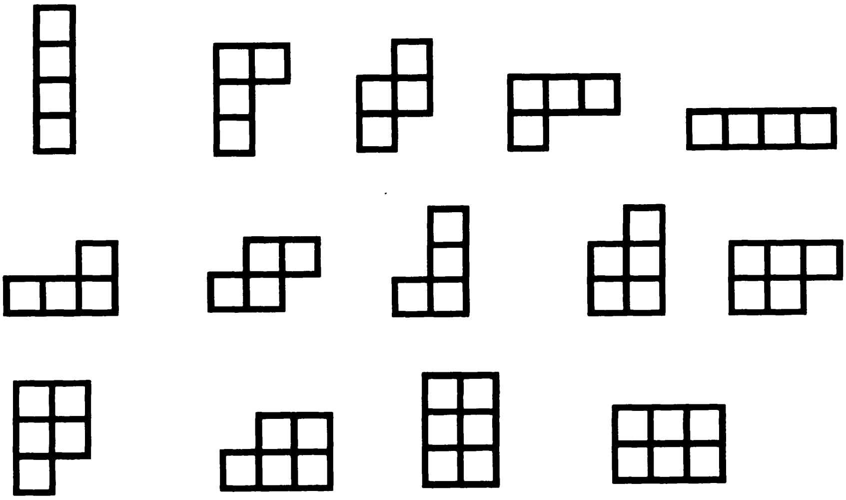


Figure 2

The Catalan numbers are ubiquitous in mathematics, showing up in the enumeration of lattice paths, planar trees, ballot sequences, fluctuations of the lead, Young tableaux, stack sortable permutations, triangulations of an n-gon, increasing functions, binary trees, and so on. In the last few years they have become a standard item in combinatorics texts, so we will resist the temptation to go into detail, and refer instead to Comtet [**3**], Wilf [**13**], Stanley (especially the forthcoming volume 2) [**12**], and the Schaum's outline by Balakrishnan [**1**], among many others. The entertaining survey by Gardner [**5**] is highly recommended.

The Catalan numbers are defined recursively by $C_0 = 1$ and $C_{n+1} = C_0C_n + C_1C_{n-1} + \cdots + C_nC_0$. The first few terms are $1, 1, 2, 5, 14, 42, 132, 429, \ldots,$ the general term is

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

The generating function for the Catalan numbers is $C(z) := \sum_{n=0}^{\infty} C_n z^n = (1 - \sqrt{1-4z})/(2z)$. We need the following three facts:

$$C(z) = 1 + z(C(z))^2, \quad \text{or} \quad C = 1 + zC^2 \tag{1}$$

$$C(z)\sqrt{1-4z} = 1 - zC^2(z), \quad \text{or} \quad C\sqrt{1-4z} = 1 - zC^2 \tag{2}$$

and

$$C^2\sqrt{1-4z} = 1 - z^2C^4. \tag{3}$$

The first follows from the definition of C_n, the third results from multiplying the first two, and a combinatorial proof of the second is included at the end of the article.

If we sketch the first few cases we see that there are 1, 2, 5, and 14 path pairs of lengths 2, 3, 4, and 5, respectively. Generalizing a bit, we let $b(n, k)$ be the number of nonintersecting paths from the origin that are k apart after n steps. Call these *partial path pairs*. (The actual distance between the endpoints is $k\sqrt{2}$, but we never need the $\sqrt{2}$ and omit it.) Table 1 illustrates the first few values of the $b(n, k)$.

TABLE 1

$n \backslash k$	1	2	3	4	5	6	7
1	1						
2	2	1					
3	5	4	1				
4	14	14	6	1			
5	42	48	27	8	1		
6	132	165	110	44	10	1	
7	429	572	429	208	65	12	1

The next observation is that $b(n+1, k) = b(n, k-1) + 2b(n, k) + b(n, k+1)$ for $k \geq 2$. This holds since the two paths can get closer at the last step in one way (the lower path goes north while the upper path goes east), they can stay the same distance apart in two ways (both east or both north), or they can get further apart in one way (lower path east, upper path north). This holds even for $k = 1$ if the nonintersecting condition is interpreted as $b(n, 0) = 0$ for $n \geq 1$.

Now an easy induction gives us

$$b(n,k) = \frac{k}{n}\binom{2n}{n-k}.$$

Since partial path pairs that are one apart after $n - 1$ steps can be completed to path pairs in n steps, we see that the number of path pairs of length n is

$$b(n-1,1) = \frac{1}{(n-1)}\binom{2(n-1)}{(n-1)-1} = \frac{(2n-2)!}{n!(n-1)!} = C_{n-1}.$$

With the preliminaries out of the way we can proceed to the proof of the main theorem. The generating function for the first column of Table 1 is

$$\sum_{n=1}^{\infty} C_n z^n = C(z) - 1 = zC^2(z).$$

The generating function for the k^{th} column is

$$\sum_{n=0}^{\infty} b(n,k)z^n = \left(zC^2(z)\right)^k = \sum_{n=k}^{\infty} \frac{k}{n}\binom{2n}{n-k}z^n,$$

as can be seen by first looking at the last place where the two paths are one unit apart, from there continuing until the last time the two paths are two units apart, and so on. The sequence of steps that takes the paths from m units apart (for the last time) to $m + 1$ units apart is the same for all m.

If we square each of these generating functions, we find that

$$\left(\sum_{n=0}^{\infty} b(n,m)z^n\right)^2 = \left(zC^2(z)\right)^{2m} = \sum_{n=2m}^{\infty} \frac{2m}{n}\binom{2n}{n-2m}z^n.$$

Thus if we define

$$B_m^*(z) := \left(\sum_{n=0}^{\infty} b(n,m)z^n\right)^2 = \sum_{n=2m}^{\infty} B^*(m,n)z^n$$

then

$$B^*(m,n) = \sum_{j=0}^{n} b(j,m)b(n-j,m) = \frac{2m}{n}\binom{2n}{n-2m}.$$

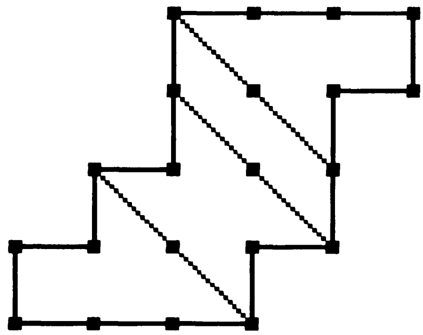


Figure 3

The combinatorial meaning of $B^*(m, n)$ is the weighted number of path pairs of length n, where the weights are the number of times that the two paths are m units apart. Thus, $B_m^*(z) = \sum_{n=2m}^{\infty} B^*(m, n)z^n$ can be thought of as the generating function for the shopkeeper counting the number of m dollar bills. Similarly, $mB^*(m, n)$ is the area of all the slices consisting of m squares on a diagonal between two paths of a path pair of length n (or, more financially, as the amount of money coming in as m dollar bills). Here is an illustration of a typical path pair contributing three slices of length 2 to the count $B^*(2, 9)$.

If we mark each square involved in each of these slices, then this path pair contributes six squares, as illustrated in Figure 4.

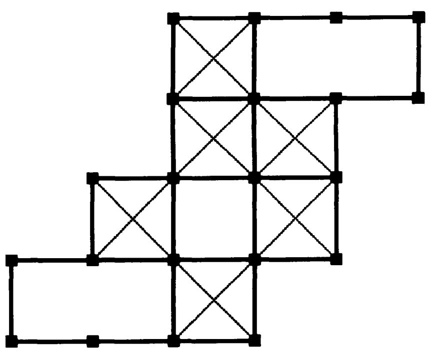


Figure 4

The main theorem will be proven if we can show that

$$\sum_{m=1}^{n} mB^*(m, n) = 4^{n-2}.$$

This can be done by manipulation of binomial coefficients (see [7, p. 87] for a similar result and similar hand waving), but we proceed by generating functions. If

we tabulate the first few values of the $B^*(m, n)$, we have

TABLE 2

$n \setminus m$	1	2	3
0	0	0	0
1	0	0	0
2	1	0	0
3	4	0	0
4	14	1	0
5	48	8	0
6	165	44	1
7	572	208	12

and the column generating functions are $z^2C^4, z^4C^8, z^6C^{12}, \ldots$ while the generating function for the sequence $1, 2, 3, 4, \ldots$ is $1/(1-z)^2$. Hence the generating function that we want is

$$z^2C^4 + 2(z^2C^4)^2 + 3(z^2C^4)^3 + \cdots = z^2C^4 \cdot \frac{1}{(1 - z^2C^4)^2} = \frac{z^2C^4}{(C^2\sqrt{1-4z})^2}$$

$$= \frac{z^2}{1-4z} = \sum_{n=2}^{\infty} 4^{n-2}z^n.$$

The second equality follows from equation (3). This completes the proof of the main theorem except for a proof of (2). We start by rewriting (2) as

$$\frac{1}{\sqrt{1-4z}} = \frac{C}{1 - zC^2} = C + C(zC^2) + C(zC^2)^2 + \cdots.$$

Consider now a single path starting at the origin, with unit steps east and north, ending up on the line $y = x$ after $2n$ steps. The number of such paths is $\binom{2n}{n}$ and thus the appropriate generating function is

$$\sum_{n=0}^{\infty} \binom{2n}{n} z^n = \frac{1}{\sqrt{1-4z}}.$$

We decompose each such path into subpaths with a new subpath starting whenever the original path crosses the diagonal line $y = x$. The generating function for the number of paths that start and end on the diagonal but never cross is C. If we want to eliminate degenerate paths, the appropriate generating function is $C - 1 = zC^2$. We take any one of these binomial paths and assume momentarily that it starts off below the diagonal. We would pick up a factor C until the first time it crosses the line $y = x$. A crossing implies that we have a nontrivial subpath on the other side of $x = y$ and this continues for each crossing. The case in which we start by going above the diagonal corresponds to multiplying by the constant term 1 in the first factor $C = 1 + z + 2z^2 + 5z^3 + \cdots$, thus allowing for the possibility that the initial subpath below the diagonal has no edges.

As with almost anything involving binomial coefficients, there is a large body of related literature. One good source that collects and extends many related results is the monograph of Bousquet-Mélou [**2**]. In [**8**] there is another moment counting problem involving a different Catalan setting, but where the final answer is a very similar 4^{n-1}. Yet another 4^{n-1} result in a Catalan setting is in [**6**]. This turns out to be analogous to our present result and indeed computes the higher moments as

well, but the proof is less combinatorial and more involved with the Eulerian and tangent numbers coming into play as coefficients.

ACKNOWLEDGMENTS. We would like to thank the Howard University Combinatorics group for their encouragement and support.

REFERENCES

1. V.K. Balakrishnan, *Theory and Problems of Combinatorics*, Schaum's Outline Series, McGraw-Hill, New York, 1995.
2. M. Bousquet-Mélou. *q-Énumération de polyonoms convexes*, Publications du LaCIM, (9) Univ. du Québec à Montréal, 1991.
3. L. Comtet, *Advanced Combinatorics*, Reidel publishers, Boston, 1974.
4. J. Fürlinger and J. Hofbauer, q-Catalan numbers, *J. Combinatorial Theory A*. **40** (1985) 248–264.
5. M. Gardner, Mathematical games column, *Scientific American*, June 1976. Also reprinted and updated in *Time Travel*, Freeman, New York, 1988.
6. S. Getu, L.W. Shapiro, and W-J Woan, Runs, slides, and moments, *SIAM J. of Algebraic and Discrete Methods* **4** (1983) 459–466.
7. D. Knuth, *The Art of Computer Programming, vol. 3: Sorting and Searching*, Addison-Wesley, Redwood City, 1973.
8. G. Kreweras, Aires des chemins surdiagonal et application a un problème économique, *Cahiers du Bulletin Universitaire de Recheré Operationnelle* **24** (1976) 1–8.
9. J. Levine, Note on the number of pairs of non-intersecting routes, *Scripta Mathematica* **24** (1959) 335–338.
10. G. Pólya, On the number of certain lattice polygons, *J. Comb. Th.* **6** (1969) 102–105.
11. L.W. Shapiro, A Catalan triangle, *Discrete Mathematics* **14** (1976) 83–90.
12. R.P. Stanley, *Enumerative Combinatorics*, Cambridge University Press, Cambridge, v.1 (1986), v.2 (to appear).
13. H.S. Wilf, *Generatingfunctionology*, Academic Press, San Diego, 1992.

Wen-Jin Woan and Lou Shapiro
Department of Mathematics,
Howard University
Washington DC, 20059
wjw@scs.howard.edu
lws@scs.howard.edu

Douglas G. Rogers
Fernley House, The Green
Croxley Green, United Kingdom, WD3 3HT
drogers@cs.bgsu.edu

Good Matrices: Matrices that Preserve Ideals

R. Bruce Richter and William P. Wardlaw

1. INTRODUCTION. The following problem of Mawaffaq Hajja appeared in the Monthly [**5**]:

> Let A and B be matrices with integer entries of sizes r by n and n by r, respectively, with $r < n$. Suppose that AB is an r by r identity matrix. Show that A can be enlarged to an n by n integral matrix having an integral inverse.

The case $r = 1$ is fairly well-known: if $A = [a_1, \dots, a_n]$ and $B = [b_1, \dots, b_n]^T$ are such that $AB = [1]$, then $a_1, \dots, a_n$ are relatively prime. In this case, there is an $n \times n$ integral matrix that has an integral inverse and whose first row is $a_1, \dots, a_n$; see [**10**, Thm. II.1, p. 13].

Hajja's Problem suggests a natural generalization that has not received much attention and will be a main topic of this article: what are the properties that two or more given rows must have so they can serve as the first rows of an invertible matrix?

The property central to our paper is the following. A matrix A with entries in a communtative ring R with unity is *left good* if, for every vector $\mathbf{x}$, the ideal $(\mathbf{x}A)$ generated by the entries in the vector $\mathbf{x}A$ is the same as the ideal $(\mathbf{x})$ generated by the entries in the vector $\mathbf{x}$. In the context of matrices with integral entries, this is equivalent to requiring that the greatest common divisor of the entries in $\mathbf{x}A$ is the same as the greatest common divisor of the entries in $\mathbf{x}$. Since, for any matrix A and any vector $\mathbf{x}$, it is obvious that $(\mathbf{x}A) \subseteq (\mathbf{x})$, the content of left goodness is in the reverse containment. Our goal is to prove the following.

Main Theorem. *Consider the following statements about an $r \times n$ matrix A over the commutative ring with unity R.*

(1) *The rows of A extend to a basis of $R^{1\times n}$.*
(2) *A can be enlarged to an $n \times n$ matrix invertible over R.*
(3) *A has a right inverse over R.*
(4) *The ideal generated by all $r \times r$ subdeterminants of A is R.*
(5) *A is left good.*

Then (i) (1) $\Leftrightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (4) $\Leftrightarrow$ (5).
(ii) *If R is a principal ideal ring, then each of* (1)–(5) *is equivalent to*

(6) *A has Smith Normal Form $[I_r\ 0]$.*

Hajja's problem is solved by observing that, over the integers, (3) implies (2).

We note that there is another direction in which the $r = 1$ case of Hajja's problem can be generalized. From the Main Theorem we see that if $\mathbf{a}$ is a vector with entries in a principal ideal ring R (such as the integers or the ring $k[x]$ of

polynomials in one variable over a field k [**6**, p. 180, ex. 9]) such that $(\mathbf{a}) = R$, then $\mathbf{a}$ is the first row of an invertible matrix over R. This statement has been made quite famous by the proof that the result also holds over a polynomial ring $k[x_1, \dots, x_d]$ in several variables over a field. This fact is crucial in the demonstration of Serre's Conjecture that every projective module over $k[x_1, \dots, x_d]$ is free [**8**, pp. 488-492]. As far as we know, no one seems to have looked for a characterization of those rings for which the $r = 1$ case holds.

To allay the reader's suspense, we include the following solution to Hajja's problem, offered to us by a referee. This is essentially the published solution [**5**].

Proposition 1. *Let A be an $r \times n$ matrix with integral entries. Let B be any integral matrix such that $AB = I_r$. Then there exist integral matrices A' and B' such that*

$$\begin{bmatrix} A \\ A' \end{bmatrix} \begin{bmatrix} B & B' \end{bmatrix} = I_n.$$

Proof: Let $A_1, \dots, A_r$ be the rows of A and let $B_1, \dots, B_r$ be the columns of B. We will add A_{r+1} to A to get A' and B_{r+1} to B to get B' such that $A'B' = I_{r+1}$. For A_{r+1}, choose any vector in the left null space of B having relatively prime coordinates.

Now let C be any column such that $A_{r+1}C = 1$. We consider choices for B_{r+1} of the form

$$B_{r+1} = C + \alpha_1 B_1 + \cdots + \alpha_r B_r.$$

For $i = 1, 2, \dots, r$ the relation $A_i B_{r+1} = 0$ is equivalent to $\alpha_i = -A_i C$. Hence $B_{r+1} = C - (A_1 C)B_1 - \cdots - (A_r C)B_r$ is a suitable column by which to extend B. ∎

2. AN INTRODUCTION TO GOOD MATRICES. In this section, we provide some examples to illustrate good matrices and prove some elementary facts about them.

Consider the integral matrix

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

If $\mathbf{x} = [x_1\ x_2]$ is any integral vector, then $\mathbf{x}A = [x_1\ x_1 + x_2]$. If $z = ax_1 + bx_2$ is an integral combination of x_1 and x_2, then $z = (a-b)x_1 + b(x_1 + x_2)$ is also an integral combination of x_1 and $x_1 + x_2$. Therefore, the ideal generated by the entries in $\mathbf{x}$ is contained in the ideal generated by the entries in $\mathbf{x}A$. As mentioned earlier, the reverse inclusion always holds and, therefore, this matrix A is left good.

Consider next the integral matrix

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

This matrix is not a good matrix since the ideal generated by $\mathbf{x} = [1\,1]$ is $\mathbb{Z}$, while the ideal generated by $\mathbf{x}A = [0\,2]$ is the set of even integers.

The following elementary properties of left good matrices are useful ones for us.

Lemma 2. *Let A and B be matrices with entries in the commutative ring R.*

(1) *If both A and B are left good and AB is defined, then AB is left good.*
(2) *If AB is left good, then A is left good.*

(3) *If $AB = I$, then A is left good.*
(4) *If P and Q are invertible over R and PAQ is defined, then PAQ is left good if and only if A is left good.*

The reader will note that (3) is the implication (3) $\Rightarrow$ (5) of the Main Theorem.

Proof:

(1) For any vector $\mathbf{x}$ for which $\mathbf{x}A$ is defined, $(\mathbf{x}) = (\mathbf{x}A) = ([\mathbf{x}A]B) = (\mathbf{x}[AB])$.
(2) Since AB is left good, for any vector $\mathbf{x}$, $(\mathbf{x}) = (\mathbf{x}AB)$. Since $(\mathbf{x}AB) \subseteq (\mathbf{x}A) \subseteq (\mathbf{x})$ equality holds throughout and $(\mathbf{x}) = (\mathbf{x}A)$, as required.
(3) The matrix I is trivially left good, so AB is left good. By (2), A is left good.
(4) Immediate from (1) and (3). ∎

We conclude this section with the relationship between being left good and the subdeterminants of the matrix, i.e., the equivalence of (4) and (5) in the Main Theorem.

Proposition 3. *Let A be an $r \times n$ matrix with entries in a commutative ring R. Then A is left good if and only if the ideal D_r generated by the $r \times r$ subdeterminants of A is R.*

Proof: Suppose A is left good and suppose D_r is a proper ideal in R. Let $\bar{R}$ be the quotient ring R/D_r. Let $\bar{A}$ be the matrix with entries in $\bar{R}$ corresponding to A. All the $r \times r$ subdeterminants of $\bar{A}$ are 0 in $\bar{R}$, so the rank of $\bar{A}$ is less than r.

Therefore, there is a nonzero vector $\bar{\mathbf{x}}$ with entries in $\bar{R}$ such that $\bar{\mathbf{x}}\bar{A} = \mathbf{0}$; this is a simple consequence of [**9**, Thm. 51, p. 159]. Let $\mathbf{x}$ be a vector with entries in R corresponding to $\bar{\mathbf{x}}$. Since A is left good, $(\mathbf{x}) = (\mathbf{x}A) \subset D_r$, whence $\bar{\mathbf{x}} = \mathbf{0}$ in $\bar{R}$, a contradiction.

For the converse, suppose $D_r = R$ but A is not left good. Let $\mathbf{x}$ be a vector with entries in R such that $(\mathbf{x}) \neq (\mathbf{x}A)$. Since $(\mathbf{x}A) \subsetneq (\mathbf{x})$, $(\mathbf{x}A) \subsetneq R$. In the quotient $\bar{R} = R/(\mathbf{x}A)$, the vector $\bar{\mathbf{x}}$ corresponding to $\mathbf{x}$ is not $\mathbf{0}$, but $\bar{\mathbf{x}}\bar{A}$ is $\mathbf{0}$.

Then $\bar{A}$ has rank less than r, so there is a nonzero annihilator $\bar{a}$ of the $r \times r$ subdeterminants of $\bar{A}$. (We heartily recommend the treatment of the rank of a matrix given in [**3**, Ch. 4] or [**9**, Ch. VIII].) Thus, in R, $a \notin (\mathbf{x}A)$ and $aD_r \subseteq (\mathbf{x}A)$. This contradicts the assumption that $D_r = R$, since then $a \in aD_r$. ∎

3. THE SITUATION FOR COMMUTATIVE RINGS WITH UNITY.

In this section, we complete the proof of (i) of the Main Theroem. We have yet to prove the equivalence of (1) and (2) and the implication (2) implies (3). The former equivalence is contained in the following fact, which is given in [**11**, Cor. 1.2}.

Lemma 4. *Let R be a commutative ring with unity and let P be an $n \times n$ matrix with entries in R. Then the following are equivalent:*

(1) *The rows of P form a basis of $R^{1 \times n}$.*
(2) *P has an inverse with entries in R.*
(3) det *P is a unit in R.*

We remind the reader that an element a in a commutative ring R with unity 1 is a *unit* if there is an element b of R such that $ab = 1$.

(2) implies (3). Suppose A extends to the $n \times n$ matrix

$$A^* = \begin{bmatrix} A \\ A' \end{bmatrix}$$

with A^* invertible over R. Let B^* be an inverse of A^* and write $B^* = [B\,B']$, with B an $n \times r$ matrix. Clearly,

$$A^*B^* = \begin{bmatrix} AB & AB' \\ A'B & A'B' \end{bmatrix} = I_n,$$

which implies that $AB = I_r$, as required.

This completes the proof of (i) of the Main Theorem. ∎

4. PRINCIPAL IDEAL RINGS. A *principal ideal ring* is a commutative ring R with unity in which every ideal is principal; i.e., if J is an ideal, then there is an element $a \in R$ such that $J = (a)$. We begin with a discussion of the Smith Normal Form of a matrix with entries in the principal ideal ring R.

Let U_k denote the set of all $k \times k$ matrices invertible over R. Two $r \times n$ matrices A and B over R are *equivalent*, written $A \sim B$, if there is a $P \in U_r$ and a $Q \in U_n$ such that $B = PAQ$. It is easy to see that $\sim$ is an equivalence relation.

Let A be an $r \times n$ matrix with entries in the principal ideal ring R. We assume for the moment that $r \leq n$. A *Smith Normal Form* of A is an $r \times n$ matrix

$$S(A) = \begin{bmatrix} d_1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & d_3 & \cdots & 0 & 0 & \cdots & 0 \\ & & \vdots & & \vdots & & & \\ 0 & 0 & 0 & \cdots & d_r & 0 & \cdots & 0 \end{bmatrix} \tag{1}$$

such that $A \sim S(A)$ and $(d_r) \subseteq (d_{r-1}) \subseteq \cdots \subseteq (d_1)$.

If it happens that A is an $n \times r$ matrix with $r < n$, then the Smith Normal Form for A is the transpose of the matrix displayed in (1) and is still denoted by $S(A)$.

A matrix A can have many different Smith Normal Forms. For example, we can replace any d_i by $d_i u$, for any unit u in R. Over the integers, we usually require $d_i \geq 0$.

Two elements a and b of a commutative ring R are *associates* if there is a unit u of R such that $a = ub$. Over general commutative rings, we can have $(a) = (b)$ without a and b being associates. (See [**1**], [**2**], or [**3**, Ex. 5.15, p. 42].) The following result is proved in [**3**, Thm. 15.24, p. 194].

Theorem 5. *Let R be a principal ideal ring and let A be any matrix with entries in R. Then*:

(1) *A has a Smith Normal Form*;
(2) *Suppose $d_1, d_2, \ldots, d_r$ and $d_1', d_2', \ldots, d_r'$ are the diagonal entries of two $r \times n$ matrices that are otherwise zero and suppose the d_i are the entries of a Smith Normal Form of A. Then the d_i' are the entries of a Smith Normal Form for A if and only if, for each $i = 1, 2, \ldots, r$, d_i' is an associate of d_i*;
(3) *With $d_1, d_2, \ldots, d_r$ as in* (2), *the ideal generated by the $i \times i$ subdeterminants of A is $(d_1 d_2 \ldots d_i)$.*

We remark that Theorem 5 is trivial if R is a field, while the proof when $R = \mathbb{Z}$ is aided by the Euclidean algorithm. Basically, one wants to do reversible row and column operations to turn A into $S(A)$. For an arbitrary principal ideal ring, the proof is somewhat technical.

We are now prepared for the proof of (ii) of the Main Theorem.

Proof of MT (ii) It suffices to prove that (4) implies (6) and (6) implies (2).

(4) implies (6). The ideal generated by the $i \times i$ subdeterminants of A is $(d_1 d_2 \dots d_i)$. We are assuming that $(d_1 d_2 \dots d_r) = R$, which means each d_i is a unit. Hence, the diagonal entries of a Smith Normal Form for A are all units, i.e., the matrix $[I_r\, 0]$ is a Smith Normal Form for A.

(6) implies (2). The hypothesis is that $A = P[I_r\, 0]Q$, for some $P \in U_r$ and $Q \in U_n$. Then

$$A^* = \begin{bmatrix} P & 0 \\ 0 & I_{n-r} \end{bmatrix} Q$$

is an extension of A to an $n \times n$ matrix A^* with inverse

$$Q^{-1} \begin{bmatrix} P^{-1} & 0 \\ 0 & I_{n-r} \end{bmatrix}. \qquad \blacksquare$$

The proof of (6) $\Rightarrow$ (2) works over any commutative ring R with unity, while (4) $\Rightarrow$ (6) depends only on the existence of a Smith Normal Form for A. Thus, these implications do not depend directly on the fact that the ring is a principal ideal ring. However, matrices over general commutative rings do not necessarily have Smith Normal Forms. This is where we use the assumption that R is a principal ideal ring.

5. RELATED MATTERS. In this section, we deal with items that are closely related to the Main Theorem but are not required for its proof.

The proof of Proposition 1 requires only the following property of the ring R: for any $r < n$ and any given $r \times n$ matrix A with a right inverse, there is some vector $\mathbf{v}$ such that $A\mathbf{v} = \mathbf{0}$ and $(\mathbf{v}) = R$. Thus, this condition on R is sufficient for the implication (3) $\Rightarrow$ (2).

Conversely, if (3) $\Rightarrow$ (2) always holds over R, then we claim that for any $r < n$ and any given $r \times n$ matrix A with a right inverse, there is some vector $\mathbf{v}$ satisfying $A\mathbf{v} = \mathbf{0}$ and $(\mathbf{v}) = R$. To see this, note that A extends to a matrix

$$\begin{bmatrix} A \\ A' \end{bmatrix}$$

that is invertible, with inverse $\begin{bmatrix} B & B' \end{bmatrix}$. If $\mathbf{v}$ is any column of B', then $A\mathbf{v} = \mathbf{0}$. On the other hand, there is a row $\mathbf{u}$ of A' such that $\mathbf{uv} = 1$, so $(\mathbf{v}) = R$.

More generally, for principal ideal rings we have the following fact, which we need later.

Proposition 6. *If $r < n$ and A is any $r \times n$ matrix with entries in a principal ideal ring R, then there is a vector $\mathbf{v}$ such that $A\mathbf{v} = \mathbf{0}$ and $(\mathbf{v}) = R$.*

Proof: By Theorem 5, A has a Smith Normal Form, so there are invertible matrices P and Q and an $r \times r$ diagonal matrix D such that

$$PAQ = [D\, 0].$$

Thus, $AQ = [P^{-1}D\, 0]$, so we may take the last column of Q for $\mathbf{v}$. $\blacksquare$

The case $r = n$ of the Main Theorem is quite interesting in its own right, since in this case all six of the statements are equivalent. In particular, we note the following (see [**12**]).

Proposition 7. *Let R be a commutative ring with unity and let A be an $n \times n$ matrix over R. Then A is left good if and only if A is invertible over R.*

Proof: By the equivalence of (4) and (5) in the Main Theorem, A is left good if and only if $(\det A) = R$. Clearly, $(\det \mathrm{A}) = R$ if and only if $\det A$ is a unit. By Lemma 4, $\det A$ is a unit if and only if A is invertible. ■

The same referee who in Proposition 1 improved our solution of Hajja's problem offered the following observation. Suppose that, over any commutative ring with unity, A and B^T are $(n-1) \times n$ matrices such that $AB = I_{n-1}$. Then A and B extend to $n \times n$ matrices A' and B' such that $A'B' = I_n$. For example, in the case $n = 3$ and $r = 2$, let A_1 and A_2 be the rows of A and let B_1 and B_2 be the columns of B. Then $A_i B_j$ is 1 if $i = j$ and is 0 otherwise, i.e., $A_i B_j$ is the Kronecker delta δ_{ij}. The additional third row A_3 to add to A must be orthogonal to both B_1 and B_2, which virtually forces the choice $A_3 = B_1 \times B_2$. Similarly $B_3 = A_1 \times A_2$. That $A_3 B_3 = 1$ follows from the relation ([**4**], ex. 9, p. 425)

$$(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) = (\mathbf{a} \cdot \mathbf{c})(\mathbf{b} \cdot \mathbf{d}) - (\mathbf{a} \cdot \mathbf{d})(\mathbf{b} \cdot \mathbf{c}).$$

For any $n \geq 2$, define the *generalized cross product* $\otimes(\mathbf{u}_1, \ldots, \mathbf{u}_{n-1})$ of $n-1$ vectors $\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}$ to be the formal $n \times n$ determinant whose first row is the formal vector $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ and whose remaining rows are $\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}$, where $\mathbf{e}_i$ is the standard unit vector of all 0's except for a 1 in the i^{th} position. We need the following (relatively straightforward) properties of $\otimes$:

(1) for each $i = 1, 2, \ldots, n-1$, $\mathbf{u}_i \cdot \otimes(\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}) = 0$; and
(2) $\otimes(\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}) \cdot \otimes(\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}) = \det(\mathbf{u}_i \cdot \mathbf{v}_j)$.

Now suppose A and B^T are $(n-1) \times n$ matrices with $AB = I_{n-1}$. Let $A_1, \ldots, A_{n-1}$ be the rows of A and let $B_1, \ldots, B_{n-1}$ be the columns of B. Then we can add the row $\otimes(B_1, \ldots, B_{n-1})$ to A and the column $\otimes(A_1, \ldots, A_{n-1})$ to B to create square matrices A' and B' satisfying $A'B' = I_n$.

Thus, for any commutative ring with unity, if A and B satisfy the assumptions of Hajja's problem and $r = n - 1$, then A and B can be extended to invertible matrices.

It is straightforward to show that if A is a left good matrix, then any matrix obtained from A by deleting rows is also left good. We have not found this particularly useful. However, it would be interesting to know the answer to the following open problem: Is every left good $r \times n$ matrix over a commutative ring R equivalent over R to the matrix $[I_r\ 0]$? This would show the equivalence of all six statements in the Main Theorem for any commutative ring.

We also note the following fact, which shows that left good matrices over principal ideal rings can be "arbitrarily bad" for multiplication on the right.

Proposition 8. *Let $r < n$, let A be an $r \times n$ left good matrix over the principal ideal ring R, and let $\mathbf{y} \in R^{r \times 1}$. Then there is an $\mathbf{x} \in R^{n \times 1}$ such that $(\mathbf{x}) = R$ and $A\mathbf{x} = \mathbf{y}$.*

Proof: By the Main Theorem, there is a matrix B such that $AB = I_r$. Let $\mathbf{x}_1 = B\mathbf{y}$. Since B is right good (defined in the obvious way), we have $(\mathbf{y}) = (B\mathbf{y}) = (\mathbf{x}_1)$.

By Proposition 6, there is a $\mathbf{v}$ such that $(\mathbf{v}) = R$ and $A\mathbf{v} = 0$. Set $\mathbf{x} = \mathbf{x}_1 + \mathbf{v}$. Then $A\mathbf{x} = \mathbf{y}$. Thus, $(\mathbf{x}_1) = (\mathbf{y}) = (A\mathbf{x}) \subseteq (\mathbf{x})$. We conclude that $(\mathbf{v}) = (\mathbf{x} - \mathbf{x}_1) \subseteq (\mathbf{x})$, so $R = (\mathbf{v}) \subseteq (\mathbf{x})$, as required. ■

In conclusion, we would like to introduce one more idea. A matrix A is *weakly left good* if $(\mathbf{x}) = R$ implies $(\mathbf{x}A) = R$. Clearly, every left good matrix is weakly left good.

Proposition 9. *Suppose R is a principal ideal ring. Then A is weakly left good over R if and only if A is left good over R.*

Proof: Only one implication is nontrivial. The proof hinges on the following fact.

Lemma 10. *Suppose R is a principal ideal ring. If $\mathbf{w}$ is a row vector and $(d) = (\mathbf{w})$, then there is a vector $\mathbf{y}$ such that $\mathbf{w} = d\mathbf{y}$ and $(\mathbf{y}) = R$.*

Given this fact, suppose A is weakly left good. We must show that if $\mathbf{w}A$ is defined, then $(\mathbf{w}A) = (\mathbf{w})$. Let d be an element of R that generates $(\mathbf{w})$. Then by Lemma 10 there is a vector $\mathbf{y}$ over R such that $\mathbf{w} = d\mathbf{y}$ and $(\mathbf{y}) = R$. Because A is weakly left good, we have $(\mathbf{y}A) = (\mathbf{y})$. Clearly, $(\mathbf{w}) = (d) = dR = d(\mathbf{y}A) = (\mathbf{w}A)$, as required. ■

Proof of Lemma 10. Note that $\mathbf{w}$ has Smith Normal Form $[d\,0\ldots0]$. Thus, there are invertible matrices P and Q over R such that $\mathbf{w} = P[d\,0\ldots0]Q$. Since $P = [c]$ is 1×1, it follows that $\mathbf{w} = [d\,0\ldots0]Q'$, for the invertible matrix $Q' = cQ$. Then $\mathbf{w} = d[1\,0\ldots0]Q'$, so we may take $\mathbf{y} = [1\,0\ldots0]Q'$. ■

ACKNOWLEDGMENT. We thank the anonymous referee who gave us the proofs mentioned in the article; this person did work above and beyond the usual call of duty. The first author's work was supported in part by the Natural Sciences and Engineering Research Council of Canada. The second author's work was supported in part by the Naval Academy Research Council.

REFERENCES

1. G. Benkart et al., Principal ideals and associates, submitted to Conference Proceedings, Groups St. Andrews 1997 in Bath.
2. G. Benkart, M. Meyerson, and W. Wardlaw, Problem, 600 *College Math. J.* **28** (1997), 146.
3. W. C. Brown, *Matrices over Commutative Rings*, Dekker, New York, 1992.
4. H. P. Greenspan and D. J. Benney, *Calculus, An Introduction to Applied Mathematics*, McGraw-Hill Book Co., New York, 1973.
5. M. Hajja, Problem 10315 *Amer. Math. Monthly* **100** (1993), 589; Solution **103** (1996), 349–350.
6. N. Jacobson, *Basic Algebra I*, W.H. Freeman and Co., San Francisco, 1974.
7. D. Joyner, D. Spellman, and W. Wardlaw, Problem 10495, *Amer. Math. Monthly* **103** (1996), 74.
8. S. Lang, *Algebra* (*2nd ed.*), Addison-Wesley Pub. Co., Inc., Menlo Park, 1984.
9. N. H. McCoy, *Rings and Ideals*, Mathematical Association of America, Washington, 1965.
10. M. Newman, *Integral Matrices*, Academic Press, New York, 1972.
11. R. B. Richter and W. P. Wardlaw, Diagonalization over commutative rings, *Amer. Math. Monthly* **97** (1990), 223–227.
12. W. P. Wardlaw, Problem 10487, *Amer. Math. Monthly* **102** (1995), 929.

Department of Mathematics and Statistics
Carleton University
Ottawa Canada K1S 5B6
brichter@math.carleton.ca

Department of Mathematics
U. S. Naval Academy
Annapolis, MD 21402
wpw@nadn.navy.mil

A Converse of the Mean Value Theorem

Jingcheng Tong and Peter A. Braza

1. INTRODUCTION. The mean value theorem for differentiation has a common geometric interpretation: there is a tangent line to the curve $y = f(x)$ defined on $[a, b]$ whose slope is the same as the secant line through the endpoints $(a, f(a))$, $(b, f(b))$. But does *every* tangent line have a corresponding parallel secant line?

2. A DIFFERENT VIEW OF THE MEAN VALUE THEOREM. The mean value theorem is: If f is continuous on $[a, b]$ and differentiable on (a, b), then there is some $c \in (a, b)$ such that $f'(c) = (f(b) - f(a))/(b - a)$. Alternatively, it can be expressed geometrically as: If f is continuous on $[a, b]$ and differentiable on (a, b) then for any subinterval $(a_1, b_1) \subset [a, b]$, there is a tangent line at a certain point $(c, f(c))$ that is parallel to the secant line passing through $(a_1, f(a_1))$ and $(b_1, f(b_1))$.

We consider the converse: given a tangent line passing through $(c, f(c))$, is there a secant line passing through some points $(a_1, f(a_1))$ and $(b_1, f(b_1))$ that is parallel to it? The fundamental question we discuss is more precisely expressed as, for each $c \in (a, b)$, is there a nonempty sub-interval $(a_1, b_1) \subset [a, b]$ such that $f'(c) = (f(b_1) - f(a_1))/(b_1 - a_1)$? We consider the strong form of the converse in which $c \in (a_1, b_1)$ and the weak form, which does not require $c \in (a_1, b_1)$. Geometrically, the strong form requires the secant line with slope $f'(c)$ to intersect the graph of f on opposite sides of c, whereas the secant line need not intersect the graph of f on opposite sides of c in the weak form.

A simple example shows that the converse need not hold in either the strong or weak forms. If $f(x) = x^3$ on $[-1, 1]$, then $f'(0) = 0$ but $(f(b_1) - f(a_1))/(b_1 - a_1) = (b_1^3 - a_1^3)/(b_1 - a_1) > 0$ whenever $-1 \leq a_1 < b_1 \leq 1$. In the following theorem we give the converse theorem in both the weak and strong forms.

3. A CONVERSE THEOREM

Theorem 1. *Let $f(x)$ be a function continuous on $[a, b]$ and differentiable on (a, b) and let c be a given point in (a, b). Then*

(1) Weak Form: If $f'(c)$ is not a total extremum value on (a, b), i.e., $f'(c) \neq \sup \{f'(x) | x \in (a, b)\}$ and $f'(c) \neq \inf \{f'(x) | x \in (a, b)\}$, then there is some subinterval $(a_1, b_1) \subset (a, b)$ such that $f'(c) = (f(b_1) - f(a_1))/(b_1 - a_1)$.

(2) Strong Form: If $f'(c)$ is not a local extremum value of $f'(x)$ on (a, b) and if c is not an accumulation point of the set $A_c = \{x \in (a, b) | f'(x) = f'(c)\}$, then there is a subinterval $(a_1, b_1) \subset (a, b)$ such that $a_1 < c < b_1$ and $f'(c) = (f(b_1) - f(a_1))/(b_1 - a_1)$.

Proof:

(1) If $f'(c)$ is not a total extremum of $f'(x)$ on (a, b) then there are $c_1, c_2 \in (a, b)$ such that $f'(c_1) < f'(c) < f'(c_2)$. Since $f'(c_i) = \lim_{x, y \to c_i} (f(x) - f(y))/(x - y)$ for $i = 1, 2$, there are subintervals $(x_1, y_1), (x_2, y_2) \subset (a, b)$ such that $(f(x_1) - f(y_1))/(x_1 - y_1) < f'(c) < (f(x_2) - f(y_2))/(x_2 - y_2)$. Without loss of generality,

suppose $x_1 \le x_2$. Consider the value $K = (f(x_1) - f(y_2))/(x_1 - y_2)$. If $K = f'(c)$, we are done. We need consider only the cases $K > f'(c)$ and $K < f'(c)$.

$K > f'(c)$. Since $g(y) = (f(x_1) - f(y))/(x_1 - y)$ is continuous on (x_1, b) and $g(y_1) < f'(c)$ while $g(y_2) > f'(c)$, there is a point $\bar{y}$ between y_1 and y_2 such that $g(\bar{y}) = f'(c)$ or $(f(x_1) - f(\bar{y}))/(x_1 - \bar{y}) = f'(c)$.

$K < f'(c)$. Since $h(x) = (f(x) - f(y_2))/(x - y_2)$ is continuous on (a, y_2) and $h(x_2) > f'(c)$ while $h(x_1) < f'(c)$, there is a point $\bar{x}$ between x_1 and x_2 such that $h(\bar{x}) = f'(c)$ or $(f(\bar{x}) - f(y_2))/(\bar{x} - y_2) = f'(c)$. ∎

(2) If $f'(c)$ is not a local extremum of $f'(x)$ on (a, b) then there is a subinterval $(a_0^*, b_0^*) \subset (a, b)$ such that $c \in (a_0^*, b_0^*)$ and $f'(c)$ is not a total extreme value of $f'(x)$ on (a_0^*, b_0^*). Let $a_1^* = (a_0^* + c)/2$, $b_1^* = (b_0^* + c)/2$, $a_{i+1}^* = (a_i^* + c)/2$, $b_{i+1}^* = (b_i^* + c)/2, \cdots$ Then $a_0^* < a_1^* < \cdots < a_i^* < \cdots < c < \cdots < b_i^* < \cdots < b_1^* < b_0^*$ and $\lim_{i\to\infty} a_i^* = \lim_{i\to\infty} b_i^* = c$.

By part (1) of this theorem, for each subinterval $(a_i^*, b_i^*) \subset (a_0^*, b_0^*)$, $(i > 0)$ there are a_i, b_i such that $(a_i, b_i) \subset (a_i^*, b_i^*)$ and $f'(c) = (f(b_i) - f(a_i))/(b_i - a_i)$. If $c \in (a_i, b_i)$ for some i, the theorem is proved. Hence we suppose $c \notin (a_i, b_i)$. By the mean value theorem for $f(x)$ on $[a_i, b_i]$, there is some $c_i \in (a_i, b_i)$ such that $f'(c_i) = (f(b_i) - f(a_i))/(b_i - a_i)$. Hence $f'(c) = f'(c_i)$. Notice that since $c_i \in (a_i, b_i) \subset (a_i^*, b_i^*)$, $\lim_{i\to\infty}(b_i^* - a_i^*) = 0$, and $c_i \ne c$, these c_i cannot coincide infinitely often. This implies there is an infinite discrete sequence c_{i_k} such that $\lim_{k\to\infty} c_{i_k} = c$. This is a contradiction since it implies that c is an accumulation point of the set $A_c = \{x \in (a, b) | f'(x) = f'(c)\}$. ∎

We have considered in Theorem 1 the case in which $f'(c)$ is not an extremum of $f'(x)$. The next theorem says that if $f'(c)$ is a local extremum of $f'(x)$ on (a, b) (i.e., a total extremum on some subinterval of (a, b) containing c) then the strong form of the converse either fails on that subinterval or f is linear in a neighborhood of c and therefore satisfies the strong form. Of course when f is linear, c is an accumulation point of $A_c = \{x \in (a, b) | f'(x) = f'(c)\}$. When f is not linear in a neighborhood of an accumulation point c and $f'(c)$ is a local extremum of $f'(x)$, then f must fail the strong form.

If $f'(c)$ is a local extremum of $f'(x)$ on (a, b), an easy example shows that the weak form may still hold. If we define $f(x) = x^3$ on $[-1, 0]$ and $f(x) = 0$ on $(0, 1]$, then $f'(0) = 0$ is an extremum of $f'(x)$, $x = 0$ is an accumulation point, and $(f(\beta) - f(\alpha))/(\beta - \alpha) = 0$ for any $\alpha, \beta \in (0,1)$ with $\alpha \ne \beta$.

Theorem 2. *Let $f(x)$ be a function continuous on $[a, b]$ and differentiable on (a, b), and let c be a given point in (a, b). Then if $f'(c)$ is a local extremum value of $f'(x)$ on (a, b) (i.e., $f'(c)$ is a total extremum value of $f'(x)$ on some subinterval $(a^*, b^*) \subset (a, b)$ containing c), then f is either locally linear about c or $f'(c) \ne (f(\beta) - f(\alpha))/(\beta - \alpha)$ whenever $\alpha < c < \beta$ and $(\alpha, \beta) \subset (a^*, b^*)$.*

Proof: It suffices to consider functions f such that $f(c) = 0$ and $f'(c) = 0$ since we can replace $f(x)$ by $f(x) - (f(c) + f'(c)(x - c))$. If f is linear in a neighborhood about c, we are done. Suppose there is a subinterval $(\alpha, \beta) \subset (a^*, b^*)$ such that $\alpha < c < \beta$ and $0 = f'(c) = (f(\beta) - f(\alpha))/(\beta - \alpha)$. If $f(\beta) = f(\alpha) > 0$, by the mean value theorem applied to $[\alpha, c]$ and $[c, \beta]$, there are $c_1 \in (\alpha, c)$ and $c_2 \in (c, \beta)$ such that $f'(c_1) = (f(c) - f(\alpha))/(c - \alpha) = -f(\alpha)/(c - \alpha) < 0$ and $f'(c_2) = (f(\beta) - f(c))/(\beta - c) = f(\beta)/(\beta - c) > 0$. This contradicts the fact that $f'(c) = 0$ is a local extremum of $f'(x)$. When $f(\beta) = f(\alpha) < 0$ we get a similar contradiction. If $f(\beta) = f(\alpha) = 0$, then $f(x) \equiv 0$ on $[\alpha, \beta]$ or, since f is

continuous, there is a point $p \in (\alpha, \beta)$ satisfying $f(p) \neq 0$. Applying the mean value theorem to both $[\alpha, p]$ and $[p, \beta]$ yields values for $f'(c_1)$ and $f'(c_2)$ of opposite sign, which contradicts the condition that $f'(c)$ is a local extremum of $f'(x)$. ∎

4. EXAMPLES FOR WHICH THE CONVERSE FAILS. Theorem 1 shows that the converse to the mean value theorem may fail at extremum values of $f'(x)$ and at certain accumulation points. In this section we give an example of a function f that has a countable number of points whose tangent lines to the curve $y = f(x)$ do not have any corresponding parallel secant lines. The function f' has extremum values at $x = (2^k - 1)/2^k$ for $k = 0, 1, 2, \dots$ and $x = 1$ and has an accumulation point at $x = 1$. We also give an example of a function g that has a derivative with an accumulation point at $x = 0$ but no secant line through points on either side of $x = 0$ whose slope matches that of the tangent line at $x = 0$. The function f fails both the strong and weak forms of the converse, while g fails the strong form but not the weak form.

Define f as follows:

$$f(x) = \begin{cases} x^3, & x \in [0, 1/4] \\ 2(1/4)^3 + (x - 1/2)^3, & x \in [1/4, 5/8] \\ 2(1/4)^3 + 2(1/8)^3 + (x - 3/4)^3, & x \in [5/8, 13/16] \\ \quad\vdots & \\ 2\left[(1/4)^3 + (1/8)^3 + \cdots + \left(\dfrac{1}{2^{k+1}}\right)^3\right] + \left(x - \dfrac{2^k - 1}{2^k}\right)^3, & x \in \left[\dfrac{2^{k+1} - 3}{2^{k+1}}, \dfrac{2^{k+1} - 3}{2^{k+2}}\right] \\ \quad\vdots & \\ 2\displaystyle\sum_{k=2}^{\infty} \left(\frac{1}{2^k}\right)^3 + (x - 1)^3, & x \in [1, 2]. \end{cases}$$

The function f is continuous on $[0, 2]$ and differentiable on $(0, 2)$. It is a monotonically increasing function satisfying $f'((2^k - 1)/2^k) = 0$ for $k = 0, 1, 2, \dots$ and $f'(1) = 0$ but $(f(x) - f(y))/(x - y) > 0$ for all $x, y \in [0, 2]$ with $x \neq y$. Hence, every tangent line has zero slope at $x = (2^k - 1)/2^k$ for $k = 0, 1, 2, \dots$ and $x = 1$ but every possible secant line has a positive slope.

The function

$$g(x) = \begin{cases} x^3 \sin \dfrac{1}{x} + \dfrac{x^2}{2} & x > 0 \\ 0 & x = 0 \\ x^3 \sin \dfrac{1}{x} - \dfrac{x^2}{2} & x < 0 \end{cases}$$

satisfies $g(x) > 0$ for $0 < x \leq 1/2$ and $g(x) < 0$ for $-1/2 \leq x < 0$ and has the derivative

$$g'(x) = \begin{cases} 3x^2 \sin \dfrac{1}{x} - x \cos \dfrac{1}{x} + x & x > 0 \\ 0 & x = 0. \\ 3x^2 \sin \dfrac{1}{x} - x \cos \dfrac{1}{x} - x & x < 0 \end{cases}$$

The value $g'(0) = 0$ is not an extremum of $g'(x)$ since $g'(1/((2n+1)\pi)) = 2/((2n+1)\pi) > 0$ for $n > 0$ and $g'(-1/((2n+1-1/n^2)\pi)) = -3/(4\pi n^4) + o(1/n^4) < 0$ as $n \to \infty$. However, since $g'(-1/((2n+1)\pi)) = 0$ and $g'(1/(2n\pi)) = 0$, the value $x = 0$ is an accumulation point of the set $A_0 = \{x \in (-1/2, 1/2) | g'(x) = g'(0)\}$. There is no interval $(a, b) \subset (-1/2, 1/2)$ with $0 \in (a, b)$ such that $(g(b) - g(a))/(b-a) = g'(0) = 0$ since $g(a) < 0 < g(b)$ while $a < 0 < b$. Hence there is no secant line through points on either side of $x = 0$ with the same slope as the tangent line at $x = 0$. Note that $g'(1/((2n - 1/n^2)\pi)) = -3/(4\pi n^4) + o(1/n^4) < 0$ as $n \to \infty$ so $g'(1/(2n\pi)) = 0$ is not an extremum on $(1/((2n+1)\pi), 1/((2n-1)\pi))$ for every $n > n_0$ for some n_0. Therefore, by Theorem 1, there is some subinterval $(a_1, b_1) \subset (1/((2n+1)\pi), 1/((2n-1)\pi))$ such that $g'(1/(2n\pi)) = 0 = (g(b_1) - g(a_1))/(b_1 - a_1)$. So, in any neighborhood of $x = 0$, the function g satisfies the weak form.

5. COMMENTS. There are some open questions that are worth considering. We gave an example of a function with a countably infinite number of points at which the converse of the mean value theorem fails. Could these points be dense in the interval? Is it possible to have an uncountable number of points? Our conjecture is "yes" to the first question and "no" to the second question. However, since the converse may fail at accumulation points, we are left with doubts.

Department pf Mathematics and Statistics
University of North Florida
Jacksonville, FL 32224
jtong@gw.unf.edu pbraza@unf.edu

So, what did I get from Robert Kanigel's book on Ramanujan? I think it reinforced my belief in the universality of the language of science. Here were two totally dissimilar people both culturally and temperamentally—from two totally different backgrounds. Each knew very little about aspects of the other's personal life, and when Hardy learnt later of Ramanujan's personal problems they came as a complete surprise to him. Of course, Hardy himself was too reserved to let Ramanujan get even a whiff of his own concerns. But when they met, which they did almost every day for nearly three years, they were on exactly the same wavelength—they spoke exactly the same language—they were totally *intimate* with each other in the language of mathematics.

From a review of R. Kanigel, *The Man Who Knew Infinity: A Life of the Genius Ramanujan*, Washington Square Press, 1992.
Reviewed by R. Tandon in *Resonance*, December, 1996.
Contributed by Richard Askey, University of Wisconsin.

Primes at a (Somewhat Lengthy) Glance

Takashi Agoh, Paul Erdős, and Andrew Granville

One can tell that

$$\begin{aligned} 17 &= 2^3 + 3^2 \\ 19 &= 2^4 + 3 \\ 37 &= 5^2 + 2^2 \cdot 3 \\ 47 &= 2 \cdot 5^2 - 3 \\ 53 &= 3^2 \cdot 7 - 2 \cdot 5 \\ 97 &= 3 \cdot 5 \cdot 7 - 2^3 \end{aligned}$$

are all prime, at a glance, since we have written each $n = A \pm B$ where each prime $\leq \sqrt{n}$ divides exactly one of A and B (and thus n is coprime with every prime $\leq \sqrt{n}$). This strange procedure is thoroughly investigated in **[1]**; in general, it is quite a challenge to so write a given prime n since the product of the primes $\leq \sqrt{n}$ is around $e^{\{1+o(1)\}\sqrt{n}}$.

A similar but more complicated method to establish the primality of n goes as follows: let $p_1 = 2 < p_2 = 3 < \cdots < p_k$ be the sequence of primes $\leq \sqrt{n}$. Write n in the form

$$n = N_1 + N_2 + \cdots + N_k, \tag{1}$$

where the set of prime divisors of each integer N_j (not necessarily positive) is precisely the set of all the primes up to p_k, other than p_j. Then, for each $j = 1, 2, \ldots, k$, we have $(n, p_j) = (N_j, p_j) = 1$ (since p_j divides N_i whenever $j \neq i$), and thus n is prime. This way of determining whether n is prime leads to our title. It turns out to be fairly easy to prove that there always is a representation as in (1):

Theorem. *Given $p_1 = 2 < p_2 = 3 < \cdots < p_k$, the first k primes, and a positive integer $n \leq \left(\prod_{i=1}^k p_i\right)\left(\sum_{j=1}^k 1/p_j\right)$, free of prime factors $\leq p_k$, there exist integers $N_1, N_2, \ldots, N_k$ with*

$$n = N_1 + N_2 + \cdots + N_k, \tag{1}$$

where each $|N_j| < \prod_{i=1}^k p_i$ and the prime divisors of N_j are precisely

$$\{p_1, p_2, \ldots, p_k\} \setminus \{p_j\}.$$

In fact we shall determine all such solutions to (1) in our proof.

Proof: Let $m = \prod_{i=1}^k p_i$, and let $m_j = m/p_j$ for each j. Assume that n is an integer in the range $0 < n \leq \sum_{j=1}^k m_j$, which is coprime to m. Define α_j to be the least positive integer for which

$$n \equiv m_j \alpha_j \pmod{p_j}$$

for each j (such an α_j exists since $(m_j, p_j) = 1$). Moreover $0 < \alpha_j < p_j$ since $(n, p_j) = 1$ (because $(n, m) = 1$ and p_j divides m).

Define $N = m_1\alpha_1 + m_2\alpha_2 + \cdots + m_k\alpha_k$. Since each $\alpha_j \geq 1$, we deduce that $N \geq \sum_{j=1}^k m_j \geq n$. Also, since p_j divides m_i whenever $j \neq i$, we deduce that

$$N \equiv m_j \alpha_j \equiv n \pmod{p_j}$$

for each j, and so $N \equiv n \pmod{m}$ by the Chinese Remainder Theorem. Thus we may write $N = n + \Delta m$ for some integer $\Delta \geq 0$. We also note that since each $\alpha_j < p_j$ thus $\Delta m < N < m_1 p_1 + m_2 p_2 + \cdots + m_k p_k = km$, so that $\Delta < k$.

Now in any solution to (1), m_j divides N_j for each j, by the hypothesis, so we can write $N_j = m_j a_j$ for some integer a_j. Since p_j divides m_i, which divides N_i, whenever $j \neq i$, we deduce from (1) that

$$m_j \alpha_j \equiv n \equiv N_j = m_j a_j \pmod{p_j}.$$

Therefore $a_j \equiv \alpha_j \pmod{p_j}$ since $(m_j, p_j) = 1$.

The condition $|N_j| < m$ is equivalent to the condition $|a_j| < m/m_j = p_j$. The only integers that are $< p_j$ in absolute value, and $\equiv \alpha_j \pmod{p_j}$, are α_j itself and $\alpha_j - p_j$. Therefore $a_j = \alpha_j - \delta_j p_j$ where $\delta_j = 0$ or 1. Conversely if $a_j = \alpha_j - \delta_j p_j$ where $\delta_j = 0$ or 1, then $|a_j| < p_j$ so that $|N_j| < m$. Moreover, all of the prime divisors of a_j are then $< p_j$ and thus the prime divisors of $N_j = m_j a_j$ are a subset of $\{p_1, p_2, \ldots, p_k\} \setminus \{p_j\}$, as required by the hypothesis. Also, since $m_j a_j = m_j \alpha_j - \delta_j m$, we have

$$\begin{aligned} N_1 + N_2 + \cdots + N_k &= m_1 a_1 + m_2 a_2 + \cdots + m_k a_k \\ &= (m_1 \alpha_1 - \delta_1 m) + (m_2 \alpha_2 - \delta_2 m) + \cdots + (m_k \alpha_k - \delta_k m) \\ &= N - (\delta_1 + \delta_2 + \cdots + \delta_k) m. \end{aligned}$$

Therefore (1) holds if and only if $\delta_1 + \delta_2 + \cdots + \delta_k = \Delta$, where each $\delta_j = 0$ or 1. Since $0 \leq \Delta < k$, it is evident that there are solutions to this, and that they are given when exactly Δ of the δ_j equal 1, and the rest of the δ_j equal 0.

Example. To clarify the notation in the proof above we show how to find all solutions to (1) for the example with $n = 101$ and $k = 4$:

We have

$$p_1 = 2, \quad p_2 = 3, \quad p_3 = 5, \quad \text{and } p_4 = 7,$$

so that $m = 2 \cdot 3 \cdot 5 \cdot 7 = 210$ and

$$m_1 = 105, \quad m_2 = 70, \quad m_3 = 42, \quad \text{and } m_4 = 30.$$

From some simple modular arithmetic we determine that

$$\alpha_1 = 1, \quad \alpha_2 = 2, \quad \alpha_3 = 3, \quad \text{and } \alpha_4 = 5,$$

which leads to $N = 105 \cdot 1 + 70 \cdot 2 + 42 \cdot 3 + 30 \cdot 5 = 521$. Therefore $\Delta = (N - n)/m = (521 - 101)/210 = 2$; and thus if $a_j = \alpha_j - \delta_j m_j$ for $j = 1, 2, 3, 4$, then exactly two of the $\delta_j = 1$, the other two of the $\delta_j = 0$. This leads to $\binom{4}{2} = 6$ representations of 101 as in (1), namely:

$$\begin{aligned} 101 &= -105 + 140 + 126 - 60 = 105 - 70 + 126 - 60 = -105 - 70 + 126 + 150 \\ &= -105 + 140 - 84 + 150 = 105 + 140 - 84 - 60 = 105 - 70 - 84 + 150 \end{aligned}$$

Corollary. *Every prime $n \geq 11$ may be 'proved' to be prime by expressing it in the form (1), where $p_1 = 2 < p_2 = 3 < \cdots < p_k$ are precisely the primes up to $\sqrt{n}$, and N_j is the product of all of those primes other than p_j.*

Proof: For each prime $11 \leq n \leq 47$ we verify the result by computing an appropriate expression of the form (1):

$11 = 3 + 2^3$; $13 = 3^2 + 2^2$; $17 = 3^2 + 2^3$; $19 = 3 + 2^4$; $23 = 3^3 - 2^2$;
$29 = 3^2 \cdot 5 - 2 \cdot 5 - 2 \cdot 3$; $31 = 3 \cdot 5 + 2 \cdot 5 + 2 \cdot 3$; $37 = 3 \cdot 5 + 2 \cdot 5 + 2^2 \cdot 3$;
$41 = 3 \cdot 5 + 2^2 \cdot 5 + 2 \cdot 3$; $43 = 3 \cdot 5 + 2 \cdot 5 + 2 \cdot 3^2$; $47 = 3 \cdot 5 + 2^2 \cdot 5 + 2^2 \cdot 3$.

That such expressions exist for each prime $n \geq 53$ may be deduced directly from our Theorem, by using the following Lemma to ensure that the hypothesis of the Theorem holds when we take $p_1 = 2 < p_2 = 3 < \cdots < p_k$ to be the primes up to $\sqrt{n}$.

Lemma. *If* $n \geq 49$ *then* $\left(\prod_{p \leq \sqrt{n}} p\right)\left(\sum_{p \leq \sqrt{n}} \frac{1}{p}\right) \geq n$.

Proof: Bertrand's postulate asserts that there is a prime in the interval $(x, 2x]$ whenever $x \geq 1$. In particular, there are primes $q \in (\sqrt{n}/2, \sqrt{n}]$ and $r \in (\sqrt{n}/4, \sqrt{n}/2]$.

Therefore, if $n > 400$, then $2, 3, 5, q, r$ are distinct primes $\leq \sqrt{n}$, so that

$$\left(\prod_{p \leq \sqrt{n}} p\right)\left(\sum_{p \leq \sqrt{n}} \frac{1}{p}\right) \geq 2 \cdot 3 \cdot 5 \cdot q \cdot r \cdot \left(\frac{1}{2} + \frac{1}{3} + \frac{1}{5}\right) = 31qr > 31\frac{\sqrt{n}}{2}\frac{\sqrt{n}}{4} > n.$$

If $121 \leq n \leq 400$, then $2, 3, 5, 7, 11$ are distinct primes $\leq \sqrt{n}$, so that

$$\left(\prod_{p \leq \sqrt{n}} p\right)\left(\sum_{p \leq \sqrt{n}} \frac{1}{p}\right) \geq 2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot \left(\frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \frac{1}{11}\right) = 2927 > n.$$

If $49 \leq n \leq 120$, then $2, 3, 5, 7$ are distinct primes $\leq \sqrt{n}$, so that

$$\left(\prod_{p \leq \sqrt{n}} p\right)\left(\sum_{p \leq \sqrt{n}} \frac{1}{p}\right) \geq 2 \cdot 3 \cdot 5 \cdot 7 \cdot \left(\frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{7}\right) = 247 > n.$$

In many ways, this proof of primality seems to be entirely without merit—one needs to know all of the primes $\leq \sqrt{n}$ for it to be useful—moreover, the expression in (1) is, in practice, ridiculously long. However, it does express a proof of primality in a single, albeit unwieldy, expression.

DEDICATORY. Paul Erdős passed away on 20th September, 1996, just a few weeks after this paper was accepted for publication.

ACKNOWLEDGMENT. Thanks to Carl Pomerance and John Selfridge for some useful comments. The third author is a Presidential Faculty Fellow and is supported, in part, by the National Science Foundation.

REFERENCE

1. R. K. Guy, C. B. Lacampagne, J. L. Selfridge, Primes at a glance, *Math. Comp.* **48** (1987), 183–202.

Dept. of Math.
Science U. of Tokyo
Noda, Chiba 278, Japan
agoh@ma.noda.sut.ac.jp

Mathematical Institute
Hungarian Academy of Sciences
Reáltanoda u. 13-15,
Budapest, Hungary

Dept. of Math.
U. of Georgia
Athens, Georgia 30602, USA
andrew@math.uga.edu

The Mathematics Education Reform: Why You Should Be Concerned and What You Can Do

H. Wu

To Chih-Han Sah, in memoriam

1. INTRODUCTION. In 1962, when the New Math was still on its ascent, 75 leading mathematicians published an open letter in this MONTHLY to chide the New Math for its excesses **[11]**. It decried empty abstraction and rigid formalism, and made a strong case for learning the fundamentals of traditional mathematics: "elementary algebra, plane and solid geometry, trigonometry, analytic geometry and the calculus." It also affirmed the importance for students to be able

> to use mathematical language with some fluency, ... to find proofs and, what may be the most important activity, to recognize a mathematical concept in, or to extract it from, a given concrete situation.

Unfortunately, this forceful statement is now all but forgotten in the mathematical research community.

Thirty-five years later, we are faced with another *mathematics education reform*. In the usual way in which this term is understood, it refers to both the *K-12 mathematics education reform* led by the National Council of Teachers of Mathematics (NCTM) and the *calculus reform*. This reform once again raises questions about the values of a mathematics education, this time not by imposing empty abstractions and rigid formalisms, but by redefining what constitutes mathematics and by advocating pedagogical practices based on opinions rather than research data of large-scale studies from cognitive psychology.

The reform has the potential to change completely the undergraduate mathematics curriculum and to throttle the normal process of producing a competent corps of scientists, engineers, and mathematicians. In some institutions, this potential is already a reality.

The purpose of this article is to discuss briefly some of the salient features of the reform, explain why the stakes are so high this time around, and finally point out some possible avenues for individual and collective action by mathematicians. Real progress in changing the direction of the reform will come only when reasoned arguments are heard from the whole mathematical community.

2. SOME SPECIAL FEATURES OF THE REFORM. The reform has its merits. For example, it has replaced some of the rote-learning in the traditional curriculum by supplying motivation and heuristic arguments. It has made students aware of the normal process of doing mathematics, such as making conjectures and looking for counter-examples. It has also made mathematics more relevant to the average student by promoting the use of realistic applications in the curriculum. This section is devoted to a few areas of concern in the reform in order to furnish

a basis for discussion in the subsequent sections; see **[24]** and **[25]** for a more detailed discussion as well as a more complete list of references.

The main focus of this article is not on the reform as an abstract idea—who does not want to improve education?—but rather on its concrete manifestations in the classroom or its explicit statements concerning instructional issues. Education is not a theoretical construct, so the reform must be judged by its performance and not by its rhetoric.

The first area of concern is the cavalier manner in which the reform texts treat logical arguments; this may well be the most conspicuous deviation from previous educational practices. Whenever a justification for a mathematical statement is given in reform texts, it is not made clear as a rule whether it is a heuristic argument that is very far from a proof or even fallacious, or actually a valid proof. Bald assertions without justifications are also made without comment. At the same time, the reform prides itself on its vigorous promotion of "higher order thinking skills" and "mathematical reasoning" (e.g., [**13**, pp. 5–6], and [**10**, pp. 20–21]). This uneasy alliance between two inherently contradictory positions breeds many awkward situations. For example, a recent article **[20]** published in the official journal of the National Council of Teachers of Mathematics advocates teaching trigonometric identities such as $\sin 2x = 2\sin x \cos x$ solely by graphing each side on the screen of a graphing calculator and observing that they coincide. There is no mention of a proof. Now if the authors had said that "in addition to proving the identity $\sin 2x = 2\sin x \cos x$, using the graphing capability of a calculator can reinforce students' confidence in the abstract argument," we could have applauded them for making skillful use of technology in the service of mathematics.

Let us consider another example. On p. 208 of **[7]**, the derivation of the quotient rule $(f/g)' = (f'g - fg')/g^2$ is given as follows: Let $Q = f/g$, then $f = Qg$. Differentiate both sides, employing the product rule for the right side, and solve for Q' to get the requisite formula. In this case, the formula is obtained by making use of the differentiability of Q, which is in fact part of what must be proved in the first place. Nevertheless, the circularity of this argument is allowed to stand in the face of prior assurances to the students that this is an "informal but mathematically sound justification" [**7**, p. ix], because the prevailing thinking is that this kind of intuitive argument is just right for most beginning math students. But what has happened to higher order thinking skills in the meantime?

One of the main goals of mathematics education enunciated in **[11]** was "to find proofs." In the current reform, this goal has been challenged, sometimes implicitly (e.g., [**13**, pp. 143–145], [**16**, p. 61]), and other times explicitly (e.g., [**12**]). In [**12**, p. 562], Mumford spoke for many reformers when he questioned why, in the context of calculus reform, we should even make a judicious and minimal presentation of proofs. Why "train them in making logical deductions" at all? His main argument is that "logical deduction has no place" in the practices of the sciences, nor in the lives of the rest of the educated public. While this point of view can be discussed on many levels, I shall follow **[12]** and simply stay within the context of calculus reform. Such an argument overlooks the fact that among the students in a typical calculus course are future math majors as well as serious users of mathematics. These two groups need rigorous mathematical training, and would not be satisfied with a steady diet of "persuasive heuristics," graphic displays, and nothing else. They comprise considerably more than one percent of the calculus student population as suggested in [**12**, p. 563] (see, e.g., [**24**, p. 1536]). Unfortunately, most reform texts, notwithstanding the fact that they exclude these two groups by design, promote themselves as texts for *all* students. If they could *explicitly* make

known this exclusion, then a good deal of the present concern about the reform would instantly disappear.

The argument of [12] also presupposes a severely utilitarian educational philosophy: if something is not useful, then throw it out of the curriculum. To a certain extent, even a liberal arts education makes some concessions to this philosophy. Recently, the Berkeley mathematics department voted to adopt as official text for "soft" calculus the applied version of [7], which is one of two books vigorously promoted in [12]. On the whole, though, most universities still manage to hold the line and endeavor to imbue students with the spirit of intellectual inquiry for its own sake. Students continue to be exposed, for example, to logical deductions in mathematics and poetic expressions in literature. Shakespeare has not yet been replaced entirely by Madison Avenue. There is something to be said about this time-honored tradition.

The same utilitarian impulse is responsible for a second major area of concern in the current reform, which is the *over-emphasis* on relevance and "real world applications." The need for applications in school mathematics curriculum is beyond debate, but are we willing to embrace a curriculum in which "the mathematics truly arises out of applications [and] the units are not centered around mathematical topics but rather application areas and themes, with the mathematical topics occurring as strands throughout the unit" [**1**]? The Interactive Mathematics Program (IMP) [**8**] comes close to realizing this rather extreme viewpoint, although all other reform texts succumb to its spell to varying degrees.

Those who over-emphasize relevance in school mathematics appear to want to reclaim the attention of the sizable number of students who are turned off by mathematics, and to hone the working skills of prospective high school graduates in order to make them more employable in the high-tech industry [**5**]. Now mathematics is a cohesive discipline with a well understood internal structure. A mathematics *education* ought to cultivate students' intellectual appreciation of this structure and cohesion. Reading the NCTM Standards [**13**]–[**15**], no one would believe that mathematics is getting its proper due in the present reform. To give a rather provocative example: a student coming out of a reform curriculum would not understand why the recent proof of Fermat's Last Theorem is a landmark event in human culture.

An application-oriented curriculum *can* furnish a valid mathematics education provided enough attention is given to *mathematical closure*. Tools developed for the purpose of solving a practical problem should be put in the proper mathematical context, and abstract ideas distilled from such solutions should preferably be applied to completely different situations to demonstrate the fundamental role of abstraction in mathematics. Unfortunately, mathematical closure is hardly ever applied in the reform. When the NCTM Standards discuss the problem of finding the roots of the cubic $5x^3 - 12x^2 - 16x + 8 = 0$ in the context of Grades 9–12 [**13**, pp. 152–153], *the only expectation of the majority of the students* is that they construct an algorithm for approximating the real roots and test it on a graphic calculator. This is all. No mention to this group of the nature of the roots of polynomials (are the real roots rational?), or the existence of real roots (if the degree is odd?), or the existence of complex roots in general, etc.

To put it in a musical context, an overly utilitarian approach to mathematics education is akin to impressing Beethoven's greatness on school students by presenting him solely as the composer of the tunes for the Huntley-Brinkley Show, the Beatles' movie *Help!*, and the recent TV ad for Acura. Even if we succeed, can we take pride in such a Pyrrhic victory?

A third area of concern in the current reform stems from the fact that mathematics is a precise technical language. Students must strive to master this new language. A tendency of the reform is to move mathematics *completely* back into the arena of everyday life where ambiguity and allusiveness thrive. A loss of precision in mathematical presentations is the result. One example is the way IMP **[8]** treats mathematical concepts. Consider the standard notion of the *expected value* of a random variable. In the IMP text, this concept is used throughout the second half of a unit called "The Game of Pig" [**8**, pp. 96–186], but it first appears in a homework problem as a term commonly used in everyday language [**8**, p. 134], and is subsequently never defined in the text proper. In the Glossary at the end of the whole book, it is stated: "*Expected value*: In a game or other probability situation, the average amount gained or lost per turn in the long run." [**8**, p. 257]. Without arguing whether such a definition is usable from a student's point of view (or whether it is even correct), I simply point out that in the *Teacher's Guide* for "The Game of Pig," teachers are alerted to the introduction of this new terminology, and that they are instructed to tell the students that "*the concept of expected value is nothing new*, ... [but] the use of such complex terminology makes it easier to state complex ideas." Whatever became of the goal to teach students to "extract a mathematical concept from a given situation"? Can this goal be accomplished if, instead of carefully guiding the students to perform the "extraction," the text systematically embeds the mathematics in the vagueness and uncertainty of everyday life?

Suppression of precision also takes the form of intentionally slighting basic algorithms and formulas. An example of the former is the various methods employed to avoid teaching even the basic multiplication and division algorithms in K-4. An example of the latter, the pre-calculus text **[19]** spends two pages (pp. 209–210) discussing the relationship between the measurements of an angle in degrees and radians, but assigns the discovery of the general formula relating the two to two exercises. Given this trend, we will soon see calculus texts which compute only derivatives of x^2, x^3, and x^4, but relegate the formula for x^n to an exercise; better yet, they will compute $\int_0^1 x^5\,dx$ and $\int_5^6 \sin x\,dx$ but leave the statement and proof of the Fundamental Theorem of Calculus to an exercise. Where will this end?

The preceding concerns all have to do with curriculum, but there are others of a different nature. The foremost is the relative neglect in the K-12 reform of the issue of teacher qualification. A main cause of the dysfunctional mathematics classroom of the seventies and eighties, which eventually led to the call for reform in **[18]**, is inadequate knowledge of mathematics. In light of this, the present emphasis of NCTM on curriculum, pedagogy, and assessment methods in [**13**]–[**15**], with no commitment to a rigorous program of re-training of the teachers in the field and a strengthening of the future teachers' mathematical education, practically guarantees the continued mediocrity (if not failure) of mathematics education in our nation **[23]**.

Another concern is with the new pedagogy, which relies heavily on constructivistic instructional strategies, such as cooperative learning and the discovery method. As a theory of learning, *constructivism* holds that the acquisition of knowledge takes place only when the external input has been internalized and integrated into one's own mind. However, the current reform transforms constructivism into a theory of *instruction* ([**10**], [**13**]–[**15**]). In order to help along this mental construction, class time in reform classrooms is reserved primarily for students to re-discover or re-invent concepts and methods of solution. Furthermore, this process of

re-discovery is facilitated by the use of cooperative learning. Having students work together in small cooperative groups is a preeminent characteristic of the present reform effort in school mathematics. In such a learning environment, the teacher ceases to be "the sage on the stage" and instead serves as "a guide on the side."

While a little bit of group learning and guiding-on-the-side is good in the classroom, *too much* of this is happening in the reform classrooms to the detriment of good education. When cooperative learning rules, teachers cannot share their insights with students or warn them against pitfalls. Moreover, students cannot learn enthusiastically from their teacher in class and do the mental construction *at home*. Just how much substantive mathematics can be learned this way?

One would be quite mistaken to regard my critical comments on the pedagogical and instructional recommendations of the reform as nothing more than a mathematician's objection to facts well-established in cognitive psychology. It is a sobering experience to read the articles by Grossen **[6]** and Anderson-Reder-Simon **[2]**, which provide critical assessments of these recommendations by an educator and three cognitive psychologists, respectively. In particular, the former points out the complexity in any successful application of cooperative learning and the lack of large-scale studies to support its unrestricted applications, while the latter takes to task many of the instructional prescriptions derived from constructivism.

3. WHY IT MATTERS. The most obvious reason why school mathematics education should matter to university professors is that a continuing influx of mathematically incompetent students would decimate the university mathematics curriculum. One can look no further than the United Kingdom to have one's worst fears confirmed. If a report released by the Council of the London Mathematical Society in October, 1995, is to be trusted, then the UK is some five years ahead of us in a mathematics education reform remarkably similar to our own in its rhetoric. If our reform takes hold, then according to **[22]**, we can look forward to a generation of students with:

(i) a serious lack of essential technical facility—the ability to undertake numerical and algebraic calculation with fluency and accuracy;
(ii) a marked decline in analytic powers when faced with simple problems requiring more than one step;
(iii) a changed perception of what mathematics—in particular of the essential place within it of precision and proof.

But the worst is yet to come. For example, to the charge that the Harvard Calculus **[7]** passes students through calculus without requiring any algebraic skill, one reply was that students' symbolic manipulative skills are much weaker than they used to be, and so some symbolic manipulation should be eliminated from calculus. In the same vein, in response to the charge that students pass through reform calculus with at best a rudimentary knowledge of algebra, the comment from reformers was that we did this long before calculus reform.

Instead of trying to uphold a certain standard and help mold as-yet-unformed minds, educators simply accept deterioration in the classroom as a given. It would be only a small step to apply such a philosophy in earnest to demand a total revamping of undergraduate, and even graduate, mathematics programs in order to fit the deficiencies of the new generation. In point of fact, such suggestions have already been made. For example, **[9]** recommends that we "Change the first two years of collegiate mathematics to match the new K-12 curriculum." Not coinciden-

tally, the opening statement of the Precalculus Project of the Calculus Bridge Consortium Based at Harvard University echoes this sentiment word for word: "Given the success of the reform calculus movement, students and teachers want reformed courses both preceding and following calculus" **[4]**. The mathematics department of a major state university started to revise all its upper divisional courses in May of 1996 in order to "mesh with the aftermath of the Harvard Calculus reform."

The logic of the reform is inexorable: once the reform is entrenched in K-12, university mathematics courses will have to follow suit. The next step will be inevitably a demand for reform in the graduate program. Thus in no time at all, the burning question of the day will be whether or not proofs are allowed only in graduate courses.

We must object to the reform because it threatens to bring down the whole education system. Indeed, our students of today will be the teachers of tomorrow, so when university courses start to deteriorate our children will be taught by teachers who are mathematically worse-equipped than those of today. Then the next wave of students will perform even more poorly, and the poor performance will incite the educators to demand a second mathematics education reform. And the vicious circle will continue. Lest such worries be construed as sheer paranoia, let me quote a recent (1996) report from the organizer of a workshop for high school mathematics teachers in a Western state (who asked to remain anonymous):

> In the afternoon we started talking about the state of students' preparation for calculus and all of them said it is getting worse year by year.... The picture they painted for me was one in which [the teachers] are nearly powerless to prevent what they see as a watering down of the curriculum because administrators, untrained in mathematics, are making the decisions based on reports filled with what they describe as NCTM jargon. One teacher... predicts that there will be no calculus course in three years because no one will be ready for it.

The reform also raises a grave concern in a different context. The economic and social well-being of our nation is critically dependent on the existence of a robust corps of technicians in science and technology: the competent mathematicians, scientists, and engineers who evolve from school students gifted in science and mathematics. Because the reform favors weaker students, the top students end up being shortchanged, and the continuous supply of this technical corps is put in jeopardy. This problem is becoming so serious that it has alarmed the U.S. Department of Education. In a refreshingly straightforward document **[17]**, it offers a criticism of the reform:

> Ultimately, the drive to strengthen the education of students with outstanding talents is a drive toward excellence for all students. Education reform will be slowed if it is restricted to boosting standards for students at the bottom and middle rungs of the academic ladder. At the same time we raise the "floor" (the minimum levels of accomplishment we consider to be acceptable), we also must raise the "ceiling" (the highest academic level for which we strive).

4. WHAT MATHEMATICIANS CAN DO. The open letter **[11]** is a remarkable document of sound educational principles in mathematics education, but it appears to be the only one that demonstrates the collective concern of the

mathematical research community for school mathematics in the past half century. It has left almost no marks, because tangible results in education can be achieved only by sustained effort. The absence of such an effort by the mathematical community at large, especially the research community, has allowed the traditional K-12 curriculum and the teaching of calculus to deteriorate, thereby opening the floodgate to a multitude of educational ideas of dubious merit. The reform is the natural product of this indifference.

Professional mathematicians have an additional obligation to break this indifference and speak out against the defects of the reform, because teachers who wish to do so are under pressure to maintain a facade of compliance. As a teacher from Pennsylvania put it: "The 'other side' is making it very uncomfortable for teachers such as me, and we are dropping like flies. Whereas university professors like you can disagree with impunity, that same privilege is denied to those of us lower on the scale."

If we wish to shake off this indifference and enter into a discussion of mathematics education, then we have to enlarge our vision concerning the teaching of mathematics. We have to temporarily abandon the narrow focus of training future mathematicians and embrace the broader and more complicated issue of educating students who have diverse goals in life. We must also learn about the reality in schools where teachers are habitually overworked and have not the luxury of intellectual contemplation. Criticisms of the reform that do not take into account deviations from our normal "universe of discourse" are not likely to find a receptive audience.

In discussing the reform, we also have to be aware of the existence of the many serious defects in the generic traditional mathematics curriculum in the schools **[24]**. A return to "business as usual" would be no cure.

One last thing we need to be aware of is that, although our professional instincts compel us to insist on rigorously proving everything, there is no faster way to lose credibility as educators than to build our whole case against the reform on this one theme alone. It is far too easy, for example, to harp on the absence of ϵ-δ proofs of the basic theorems of limit and continuity in the reform calculus texts, but a pedantic insistence on rigor is by no means the best approach to the teaching of elementary mathematics. It would be more realistic to ask that only the truly basic facts be proved in beginning courses and that there be careful differentiation between what is actually proved and what is not. Gaps can always be filled later, provided no circular reasoning is involved *and* provided the students are made aware of the gaps.

What then can we do, individually and collectively? Here are a few suggestions.

The situation regarding the calculus reform is relatively simple. Since it is being carried out mainly by our peers, we should press for a vigorous debate, not only in professional journals but also in every one of our own departments. If personal and anecdotal experiences serve, most mathematicians active in research regard the teaching of calculus as something unworthy of serious attention. The time to change this attitude is *now* before the reform gets out of control.

The K-12 reform is inherently more complicated and calls for efforts in more than one direction. First of all, NCTM is currently revising its *Standards* **[13]**–**[15]** for a second edition. It has created a Commission on the Future of the Standards and has asked several mathematics organizations, such as MAA, AMS, and SIAM, to create their own committees to work closely with the Commission over the next three years. These committees are to provide sustained advice and information. We should seek out members of these committees to give them our opinions on

the reform in general and on the *Standards* in particular. This is our chance to infuse the *Standards* with more mathematical substance and a more balanced viewpoint.

In the meantime, we should offer our critical comments on the reform. In spite of pleas for the mathematical community to "speak with one voice" in support of the reform, we should keep up the fine critical tradition initiated in **[11]**. What is missing in the reform is the commitment to teach mathematics, in all its guises, without violating its integrity. If we mathematicians do not reaffirm this commitment, then who will?

Something much less easy to achieve but immensely more important is for mathematicians to help improve the training of prospective school teachers. Mathematics education on the college level is, more often than not, aimed exclusively at producing future mathematicians. The usual college mathematics courses drill the students on the technical details of fundamentals in order to prepare them for graduate work in mathematics. But for those who leave mathematics after their college degree, e.g., school teachers, such courses yield brief glimpses of the trees but never the panorama of the forest. In the words of Allyn Jackson, such an experience in mathematics is akin to "finishing a BA in English literature having done a lot of technical analysis of Shakespeare but having no idea about Shakespeare's stature in English literature." Because less than 20% of math majors go on to do graduate work, we are *addressing only 20% of our students while pretending to be teaching them all*; see **[3]** and **[23]**. A narrow focus on producing future mathematicians is a significant factor in the inadequate mathematical preparation of our school teachers.

There is no simple remedy for this educational difficulty. Larger institutions can schedule different sections of the same course to satisfy the divergent needs of the students. Smaller colleges can overcome this obstacle only by the extra dedication and ingenuity of instructors. *We are all capable of making a contribution to this important matter just by being more conscientious in carrying out our normal duties.*

A third area for possible action is direct participation. For example:

(A) Be an author of school mathematics texts.
(B) Join a group that engages in curricular activities.
(C) Act as consultant and critic on education.
(D) Work directly with one's own local school board or teachers.
(E) Speak up as a citizen and do grassroots work.

Regarding (B), the main difficulty is an almost unbridgeable chasm between educators in the K-12 reform and mathematicians, so any contribution we hope to make here requires establishing some mutual trust between the two groups. Regarding (C), despite exhortations by NSF and AMS for research mathematicians to partake of the education enterprise, there is in fact no support for critical educational writing. On the other hand, NSF funded the writing of textbooks such as *Earth Algebra* **[21]**. Life is indeed full of mysteries.

Thus far, the most effective method of making one's voice heard in K-12 education is by way of grassroots efforts. Prime examples of this are the various groups organized by parents in California, which played a substantial role in hastening the revision of the 1992 California Mathematics Framework **[10]**. These and other action groups serve the vital function of giving voice to alternative points of view and galvanizing dissent into action. If we can add our professional voices to the efforts of these groups, we can help create a potent force for change within education.

ACKNOWLEDGMENTS. In writing this article, I have received invaluable help from Richard Askey, Steven Krantz, and Henrietta Wallace. The corrections of Madge Goldman and Ralph Raimi in a preliminary version led to vast improvements in style and exposition. The suggestions of Henry Alder, Al Cuoco, Richard Escobales, Carol Gambill, Deborah Tepper Haimo, Motohico Mulase, Han Sah, John Schommer, Dick Stanley, and Lynn Steen were also very important. It gives me pleasure to thank them all.

REFERENCES

1. *ARISE*, Information material available (1993) from COMAP, Inc., Suite #210, 57 Bedford Street, Lexington, MA 02173.
2. J. R. Anderson, L. M. Reder and H. A. Simon, Applications and misapplications of cognitive psychology to mathematics education, available at *http://sands.psy.cmu.edu/personal/ja/misapplied.html*
3. R. C. Atkinson and D. Tuzin, Equilibrium in the research university, *Change*, May/June 1992, 20–31.
4. *The Precalculus Project of the Calculus Bridge Consortium Based at Harvard University*, available at *http://www.math.harvard.edu/~ calculus/Precalc/Description.html*
5. S. L. Forman and L. A. Steen, Mathematics for work, *Bulletin of the International Commission on Mathematical Instruction (ICMI)*, No. 37, Dec. 1994, 17–22.
6. Bonnie Grossen, Making research serve the profession, *American Educator*, Fall 1996, pp. 7–8, 22–27.
7. D. Hughes-Hallet et al., *Calculus*, Wiley, New York, 1994.
8. Interactive Mathematics Program, *Year* 1, Key Curriculum Press, Berkeley, CA, 1997.
9. J. J. Kaput, Long-term algebra reform: democratizing access to big ideas, in *The Algebra Initiative Colloquium, Volume I*, C. B. Lacampagne, ed., U.S. Department of Education, Washington D.C., 1995, 33–49.
10. *Mathematics Framework for California Public Schools*, California Department of Education, Sacramento, CA, 1992.
11. On the mathematical curriculum of the high school, *Amer. Math. Monthly* 69 (1962), 189–193.
12. D. Mumford, Calculus reform—For the millions, *Notices Amer. Math. Soc.* 44 (1997), 559–563.
13. *Curriculum and Evaluation Standards for School Mathematics*, National Council of Teachers of Mathematics, Reston, 1989. Available at *http:/www.enc.org/online/NCTM/280dtoc1.html*
14. *Professional Standards for Teaching Mathematics*, National Council of Teachers of Mathematics, Reston, 1991.
15. *Assessment Standards for School Mathematics*, National Council of Teachers of Mathematics, Reston, 1995.
16. *Geometry from Multiple Perspectives*, Curriculum and Evaluation Standards for School Mathematics Addenda Series, National Council of Teachers of Mathematics, Reston, 1991.
17. *National Excellence: A Case For Developing America's Talent*, U.S. Department of Education, Washington D.C., October 1993.
18. *A Nation at Risk*, U.S. Department of Education, Washington D.C., 1983.
19. The North Carolina School of Science and Mathematics, *Contemporary Precalculus Through Applications*, Janson Publications, Dedham, 1992.
20. J. Pelech and J. Parker, The graphing calculator and division of fractions, *Math. Teacher* 89 (1996), 304–305.
21. C. Schaufele and N. Zumoff, *Earth Algebra*, Preliminary Edition, HarperCollins, 1993.
22. *Tackling the Mathematical Problem*, A report to the London Mathematical Society, Institute of Mathematics and its Applications, and the Royal Statistical Society, October 1995. Available at *http://www.qmw.ac.uk/~ lms/tackling/report.html*
23. H. Wu, On the education of math majors, in *Issues in Contemporary Mathematics Instruction*, edited by E. Gavosto, S. G. Krantz, and W. G. McCallum, Cambridge University Press, 1998.
24. H. Wu, The mathematician and the mathematics education reform, *Notices Amer. Math. Society* 43 (1996), 1531–1537.
25. H. Wu, The mathematics education reform: what is it and why should you care? Available from the author or at *http://math.berkeley.edu/~ wu*

Department of Mathematics #3840
University of California
Berkeley, CA 94720-3840
wu@math.berkeley.edu

Confronting Reform

Jeremy Kilpatrick

Three times in this century, some members of the American mathematical community have attempted to reform school mathematics only to discover that others objected to the direction the reform appeared to be taking. In this paper, I sketch the issues at stake in the first two reform efforts and then turn to the third and current effort, giving particular attention to the critique offered by H. Wu [33]. The paper ends with some thoughts on the challenges of changing school mathematics.

A UNIFIED CURRICULUM. At the turn of the century, reformers at the University of Chicago High School and at several other Illinois schools were attempting to unify the secondary curriculum, principally by merging the year-long courses in algebra and geometry [25]. In 1903, E. H. Moore [15], retiring as president of the American Mathematical Society, gave a powerful impetus to the nascent reform efforts by devoting part of his presidential address to mathematics in secondary education. Moore called for "the unification of pure and applied mathematics" and "the correlation of the different subjects," to be accomplished by organizing algebra, geometry, and physics into a "thoroughly coherent four years' course."

Reaction to what was to become known as the "Chicago movement" was swift. Conservative mathematicians in the East, most prominently David Eugene Smith, although tolerant of the brash Midwesterners tinkering with new approaches, argued that the secondary classroom was a place for pure, not applied, mathematics. In particular, the mental disciplinary power of geometry, together with its aesthetic and cultural value, demanded that it be kept in a separate course.

By the time the final report of the MAA's National Committee on Mathematical Requirements [13] appeared in 1923, much of the support for a unified curriculum had shifted to Grades 7 to 9 and away from Grades 10 to 12. The report's authors wanted all students, many of whom were dropping out of school by the end of Grade 9, to have a broad view of mathematics and consequently proposed some integrated courses for the junior high school. They also suggested ways in which the curriculum of Grades 10 to 12 might be reorganized to connect algebra with geometry and to include some work in statistics and even calculus. They acknowledged, however, that although experimental unified courses were being developed, few high schools were adopting them. The movement to unify the mathematics curriculum was already fading away under attacks on mathematics as a required subject in secondary school and the growth of courses emphasizing the social uses of mathematics (primarily arithmetic). Today, the residue of the reform effort can be seen in the "general mathematics" course, the impoverished counterpart to first-year algebra.

A MODERN CURRICULUM. The next wave of reform began to build in the 1950s, as university mathematicians and school mathematics teachers joined forces to attempt to bridge what they saw as the widening gap between school and collegiate mathematics. Concerned that the "explosive development of mathemat-

ics" [**4**, p. 1] was not being reflected in the school curriculum, that too few students were entering college prepared to study advanced mathematics, and that the nation risked a serious shortage of mathematically trained personnel, reformers mounted a variety of projects to improve the teaching of school mathematics by developing new curriculum materials and retraining teachers. These efforts became known as the "new math"—"a label not so much for a cohesive set of reform proposals and activities as for an era during which a variety of reforms were undertaken" [**26**, p. 413].

Many of the reforms, but not all, were marked by an emphasis on what were seen as unifying concepts of mathematics—set, relation, function, and the like—coupled with the abstract structures—groups, rings, fields, vector spaces—into which they are organized. "Because the university mathematicians who dominated the modern mathematics movement tended to be specialists in pure rather than applied mathematics, they saw pure mathematics, with an emphasis on set theory and axiomatics, not only as the content that was missing from the school curriculum but also as providing the framework around which to reorganize that curriculum" [**26**, p. 412].

Again, the reaction was swift. Morris Kline was the first and loudest voice, arguing that aspects of the reform efforts were "wholly misguided,' 'sheer nonsense,' attempts to replace the 'fruitful and rich essence of mathematics' with sterile, peripheral, pedantic details'"[quoted in **6**, p. 55]. In the position paper "On the Mathematics Curriculum of the High School" [**16**], Kline and 64 other mathematicians offered a more measured critique—essentially arguing that anyone attempting reform needed to link school mathematics more closely to its history and to concrete applications and not to make it so abstract and formal that future nonmathematicians would be turned away. An important feature of the paper was that it offered "fundamental principles and practical guidelines." E. G. Begle [**3**], director of the School Mathematics Study Group, the largest and most prominent of the new math curriculum reform projects, expressed delight with the guidelines, claiming that most were reflected in the new textbooks, and then gently chided the authors for failing to distinguish among the different projects and their suggestions for curriculum improvement, thus effectively rejecting them all.

Once most of the new math projects had ended, Kline fired the last shot. In *Why Johnny Can't Add: The Failure of the New Math* [**12**], published in 1973, he reiterated and elaborated his opposition to the reform. Although the book was marred by a sometimes flippant tone and a persistent unwillingness to make distinctions among reform efforts, Kline offered cogent thoughts on deduction, rigor, and the language of mathematics. (Despite the book's title, it dealt with the secondary curriculum and not the teaching of arithmetic. Kline once confided that his publisher insisted on the title.) He ended the book by arguing that the appropriate direction for any reform "should be diametrically opposite to that taken by the new mathematics" [**12**, p. 144], toward mathematics as an integral part of a liberal education, with connections to culture, history, science, and other subjects. He cited with approval Moore's [**15**] call to combine mathematics with science in high school and to reduce the artificial separation between its pure and applied sides. Thus, Moore, who had pushed the earlier reform effort, was cast in opposition to the second.

The residue of the new math era may be difficult to see in today's school mathematics, but it is there. The precalculus course, for example, is a direct descendent of the elementary functions and introductory analysis courses that appeared during that time. Some of the new math's terminology and notation has

disappeared, but much survives. And various topics, such as inequalities, that the reforms introduced into school mathematics have remained. With respect to changes in the way mathematics has been taught, few of the reform proposals appear to have been extensively implemented. It is popular to declare that the new math was tried and that it failed. Studies of school practice at that time [**8**], [**17**], however, suggested that in most classrooms the reforms were never really tried.

A STANDARDS-BASED CURRICULUM. Over the past decade or so, the reform impulse has heated up once more, this time led by professional organizations under the banner of raising expectations and providing mathematical literacy for all [**19**], [**20**], [**21**], [**22**]. The publication *A Nation at Risk* [**18**] set many of the terms of the discourse: low performance on international assessments threatened the nation's economic competitiveness; declining test scores nationally meant that rigorous and measurable standards were needed. Reformers of school mathematics have argued that changes in society and within mathematics itself necessitate a more demanding school mathematics curriculum. The goal is to develop students' *mathematical power*: "Truth and beauty, utility and application frame the study of mathematics like the muses of Greek theater. Together, they define mathematical power, the objective of mathematics education" [**22**, p. 43].

Much of the leadership in promoting reform has come from the National Council of Teachers of Mathematics (NCTM). This organization, founded in 1920 to help preserve the place of mathematics in the secondary curriculum, supported but did not lead the new math reform efforts. Two decades ago, however, it began to play a more active role as a national voice for teachers, in part as a response to the widely perceived failure to change school mathematics during the new math era and in part to counter the ensuing "back to basics" backlash of the mid-1970s [**14**, pp. 22–25]. In its first, and most influential, reform document [**19**], the NCTM took the term *standard* from the rhetoric of raised expectations and accountability for results and made it a statement for judging the quality of school mathematics and for providing "an informed vision of the future" [**14**, p. 36].

The language of "mathematical power" represents an attempt to provide a vision of "what it means to be mathematically literate both in a world that relies on calculators and computers to carry out mathematical procedures and in a world where mathematics is rapidly growing and is extensively being applied in diverse fields" [**19**, p. 1]. The arguments given for reforming the school mathematics curriculum, instruction, and assessment rest on the contention that because "all industrialized countries have experienced a shift from an industrial to an information society," the mathematics that students need to know in order to be "self-fulfilled, productive citizens in the next century" [**19**, p. 3] has also changed. The changes in society have demanded that schools change as well. Although previous reform efforts had their effects, virtually all observers of U.S. school mathematics classrooms have come away convinced that change is needed.

A large part of the standards-based reform is built on the view that mathematics itself has become more computational and less formal. "In recent years, a reaction against formalism has been growing. In recent mathematical research, there is a turn toward the concrete and the applicable. In texts and treatises, there is more respect for examples, less strictness in formal exposition" [**5**, p. 344]. Even before the recent controversy over "the death of proof" [**1**], [**7**], [**10**], so-called informal geometry courses, minus proof, were being introduced into high schools as state legislatures and school districts mandated "geometry for all." For some high school teachers, the call for "decreased attention" to "Euclidean geometry as a

complete axiomatic system" and "two-column proofs" [**19**, p. 127] has been interpreted as permission to do away with proof altogether, for everyone [**14**, p. 118]. (For an eloquent defense of proof in high school geometry, together with a proposed year-long syllabus, see [**32**].) And clearly, the availability of computer software and graphing calculators has made it easier than ever before to visualize relationships and test numerous cases of a generalization before or in place of providing deductive justifications.

This time around, the negative reaction to reform proposals and activities has been slow to come. The lag may be because endorsements were sought and received for the proposals from all parts of the mathematical community. Or it may be because the proposals were framed in rather general terms, with textbooks and other materials from reform projects appearing only in the last few years. Opposition to the reform, however, has been building—much of it on the Web—for some time, and now we have an article in print by H. Wu [**33**], one of the most outspoken of the critics.

WU'S CRITIQUE. Like many before him, Wu is not concerned with making distinctions among reform proposals or between the proposals and the various activities carried out and materials developed in the name of reform. He conflates what he terms the *K-12 mathematics education reform* promoted by the NCTM [**19**], [**20**], [**21**] and the *calculus reform* stimulated largely by work of the MAA and the National Academy of Sciences [**27**]. Although these two efforts share many common features, they have rather different agendas—with one attempting to lay out a broad framework for the mathematics all American schoolchildren need to know and be able to do and the other targeting a college course seen as unsatisfactory and out-of-date.

Wu's scattershot approach to the K-12 reform relies heavily for its force on unsubstantiated claims and random anecdotes. Contending that "the reform must be judged by its performance and not by its rhetoric" [**33**, p. 947], he offers no documentation of performance whatsoever. Instead, he often uses the rhetoric of the NCTM standards documents and the content of textbooks purporting to follow the standards to support his assertions that performance must be bad.

He also hits some inappropriate targets. For example, he castigates two high school teachers [**24**] writing in the *Mathematics Teacher* for their attempt to help students see the functions involved in a trigonometric identity before establishing its validity. Quite apart from whether every article in an official journal of an organization promoting reform must reflect reform views, one can reasonably ask whether seeing the graphs of these functions alone might not help students understand the identity. And how can Wu be so certain that teachers who are having students use graphing calculators are neglecting proof just because it is not mentioned in the article?

A second example is Wu's [**33**, p. 949] disapproval of textbooks that neglect "basic formulas." He cites the precalculus textbook produced by teachers at the North Carolina School of Science and Mathematics [**2**], a book whose origins were independent of, although ultimately in harmony with, the NCTM reform. The North Carolina approach relied much more than anything NCTM has proposed on the view that *every* topic ought to be introduced with an application [**11**, p. 154]. Wu notes that in its discussion of radian measure [**2**, pp. 209–210], the North Carolina textbook fails to give the formula relating degrees and radians but instead leaves it to the exercises. The issue here, as any textbook author will recognize, is the tension between textbook as archive and textbook as tool for learning. Once a

formula is put into a text for memorization and subsequent reference, there is little point in asking the reader to find it. The omission of a formula from a textbook, reform or otherwise, should not be interpreted to imply that teachers will slight it in the classroom. Is there no value in having students develop a formula on their own? Must it always be given in the textbook? Wu is on much firmer ground elsewhere when he discusses problems associated with the use of open-ended problems [**31**]. The examples he gives clearly show that the authors of such problems have not always thought through the mathematics needed to solve them and the reasons for having students work on them.

A final example of Wu's indiscriminate approach is his criticism of the NCTM standards [**19**] for presumably arguing that information about the nature and existence of polynomial roots should be withheld from most high school students [**33**, p. 948]. The reference is to a discussion of how the solution of polynomial equations might be "differentiated in both depth and the level of formalism" [**19**, p. 152] by treating it at any of five levels. Nowhere does the document say that as a rule students should not learn about polynomial roots. What the discussion attempts is an illustration of how teachers might begin with exploration rather than abstraction, depending on the students' knowledge and experience. Wu sees this approach as advocacy of a utilitarian curriculum that never reaches "mathematical closure" (i.e., formal proof). He is right that the arguments for a standards-based curriculum are largely utilitarian, but he wrongly attributes the apparent decline in attention to proof to the standards movement alone, and he makes the same error some teachers do in interpreting a call for decreased attention to certain proof practices as sanctioning the complete elimination of formal proof itself.

The conception of the learner underlying the NCTM standards has usually been characterized as *constructivism* [**14**, p. 113], a term that has almost lost its meaning in American mathematics education. It began as the epistemological position, associated with Jean Piaget, that a learner both incorporates novel experience into existing mental structures (assimilation) and reorganizes those structures to handle more problematic experience (accommodation). Later interpreters stressed the accommodation aspect, arguing that learners actively construct knowledge rather than receiving it passively from the environment. The most radical view, which has become popular in some quarters of American mathematics education but almost nowhere else, is that the learner is an informationally closed system that cannot know an independent, pre-existing external world [**29**], [**30**].

As a theory of knowledge acquisition, constructivism says nothing about how teaching or instruction should proceed. In recent years, however, practices that encourage students to become active learners by conducting investigations, working in groups, and handling concrete objects have come to be characterized as "constructivist teaching." Only if educators all the way back to Plato—including Comenius, Pestalozzi, Herbart, Froebel, Dewey, and Montessori—were to be considered constructivists would such practices uniquely define constructivist teaching. Overblown claims have been made that radical constructivism has brought "a new revolution in mathematics education of a magnitude no less than the modern mathematics movement of the 1960s" [**28**, p. 720] and has provided the epistemology underlying the current reform, but there is no real evidence for these assertions. What is clear, however, is that the reform documents advocate some pedagogical practices that 50 years ago might have been labeled "progressive," but that today are termed "constructivist."

In his critique, Wu condemns "the new pedagogy, which relies heavily on constructivistic instructional strategies, such as cooperative learning [a form of

group work not directly connected to the standards-based reform] and the discovery method [apparently any instructional approach in which students engage in inquiry]" [**33**, p. 949]. He claims that "*too much* of this is happening in the reform classrooms to the detriment of good education" [**33**, p. 950]. It is impossible to know, in the absence of any data, what constitutes heavy reliance or "too much."

Wu cites an article [**9**] that he says laments the absence of large-scale studies supporting "unrestricted" (his term) applications of cooperative learning. Although the article does not actually make that assertion, a more critical point is that research, no matter how large in scale, can never justify the indiscriminate use of a teaching method, traditional or otherwise. Methods are always interpreted and used in different ways by different teachers for teaching different topics to different students. A new method can no more be shown universally superior than can traditional methods, whatever they might be.

The most constructive part of Wu's critique, by far, is the final section in which he urges mathematicians to become more involved in mathematics education, contributing ideas to the revision of the NCTM standards documents, helping to improve the training of prospective teachers, and participating directly in curriculum change. Many mathematicians have already been involved in the current reforms for some time, but greater participation—encouraging or critical—can only be beneficial.

To progress as a field in how we deal with efforts to improve school mathematics, however, we need not only greater participation but also a higher level of discourse about those efforts. Critiques need to be based on substantive analyses that are grounded in evidence. They should consist of more than capricious assertions and bleak prophesies. We need to move from anecdote to analysis, from evisceration to evidence, from diatribe to dialogue.

SUPPORTING ONE ANOTHER. Over the past century, the American mathematical community has become one of the most cohesive academic communities in the world. Few disciplines anywhere, for example, have organizations such as the Conference Board of the Mathematical Sciences or the Mathematical Sciences Education Board to unite all elements of the community, from elementary school teachers to applied mathematicians working in industry. At conferences bringing together selected teachers, professors, and other scholars from across the continent to discuss education problems, those concerned with mathematics are almost invariably the first to coalesce, with many already well acquainted with one another. Since the American mathematical research community emerged over a century ago [**23**], American mathematics education has benefited from a virtually continual stream of support from prominent research mathematicians who have taken an interest in education, been willing to speak out for it publicly, and helped work for educational change.

Despite the cohesiveness, however, strains appear from time to time when the school mathematics curriculum is under scrutiny. These strains can be expressed as polar opposites, although of course there is always a spectrum of opinion in between. Some favor pure mathematics; others applied. Some want mathematics taught as they learned it; others want a different approach. Some are concerned primarily with developing the next generation of mathematicians; others are concerned primarily with mathematical literacy for all. For some, the deductive side of mathematics is what counts; others prefer the empirical, fallibilist, culturally determined side.

Whenever the mathematics taught in schools seems especially removed from mathematics seen as a scientific discipline and human enterprise, the strains can become especially great. That has happened at the beginning of the century, at mid-century, and now as the century draws to a close. The tension that these disagreements entail should remind us that if we did not share so much in common, we would not have such good grounds on which to disagree and to work toward a resolution.

Change in education is notoriously complex, difficult, and unpredictable. Reform movements in mathematics education turn out neither as advocates hope nor as detractors fear. But these movements can energize those teachers who want, as Begle once put it, to teach better mathematics and to teach mathematics better. As teachers struggle to improve their practice, a reform vision can provide needed direction, and membership in a mathematical community can provide needed support.

ACKNOWLEDGMENTS. I am grateful to Alan Schoenfeld, Guershon Harel, and George Stanic for comments on an earlier version of the paper and to Mark Freitag for observations on the treatment of proof in recent textbooks.

REFERENCES

1. G. Andrews, The death of proof? Semi-rigorous mathematics? You've got to be kidding!, *Math. Intelligencer*, Fall 1994, pp. 16–18.
2. G. B. Barrett, K. G. Bartkovich, H. L. Compton, S. Davis, D. Doyle, J. A. Goebel, L. D. Gould, J. L. Graves, J. A. Lutz, and D. J. Teague, *Contemporary Precalculus Through Applications: Functions, Data Analysis and Matrices*, Janson Publications, Dedham, MA, 1992.
3. E. G. Begle, Remarks on the memorandum "On the mathematics curriculum of the high school," *Amer. Math. Monthly* 69 (1962), 425–426; see also *Math. Teacher* 55 (1962), 195–196.
4. College Entrance Examination Board, Commission on Mathematics, *Program for College Preparatory Mathematics*, CEEB, New York, 1959.
5. P. J. Davis and R. Hersh, *The Mathematical Experience*, Birkhäuser, Boston, 1981.
6. B. DeMott, The math wars, in *New Curricula*, R. W. Heath, ed., pp. 54–67, Harper and Row, New York, 1964.
7. K. Devlin, The death of proof? *Notices Amer. Math. Soc.* 40 (1993), 1352–1353.
8. J. T. Fey, Mathematics today: Perspectives from three national surveys, *Math. Teacher* 72 (1979), 490–504.
9. B. Grossen, Making research serve the profession, *Amer. Educator*, Fall 1996, pp. 7–8, 22–27.
10. J. Horgan, The death of proof, *Scientific American*, October 1993, pp. 92–103.
11. J. Kilpatrick, L. Hancock, D. S. Mewborn, and L. Stallings, Teaching and learning cross-country mathematics: A story of innovation in precalculus, in *Bold Ventures, Vol. 3: Case Studies of U.S. Innovations in Mathematics Education*, S. A. Raizen and E. D. Britton, eds., pp. 133–243, Kluwer, Dordrecht, 1996.
12. M. Kline, *Why Johnny Can't Add: The Failure of the New Math*, St. Martin's Press, New York, 1973.
13. Mathematical Association of America, National Committee on Mathematical Requirements, *The Reorganization of Mathematics in Secondary Education*, MAA, 1923.
14. D. B. McLeod, R. E. Stake, B. P. Schappelle, M. Mellissinos, and M. J. Gierl, Setting the standards: NCTM's role in the reform of mathematics education, in *Bold Ventures, Vol. 3: Case Studies of U.S. Innovations in Mathematics Education*, S. A. Raizen and E. D. Britton, eds., pp. 13–132, Kluwer, Dordrecht, 1996.
15. E. H. Moore, On the foundations of mathematics, *Science* 17 (1903), 401–416; *Bull. Amer. Math. Soc.* 9 (1903), 402–424; reprinted in *Math. Teacher* 60 (1967), 360–374.
16. On the mathematics curriculum of the high school, *Amer. Math. Monthly* 69 (1962), 189–193; *Math. Teacher* 55 (1962), 191–195.
17. National Advisory Committee on Mathematical Education, *Overview and Analysis of School Mathematics Grades K–12*, National Council of Teachers of Mathematics, Reston, VA, 1975.

18. National Commission on Excellence in Education, *A Nation at Risk*: *The Imperative for Educational Reform*, U.S. Government Printing Office, Washington, DC, 1983.
19. National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics*, Reston, VA, 1989. Available at the web site: *http://www.enc.org/reform/index.htm*
20. National Council of Teachers of Mathematics, *Professional Standards for Teaching Mathematics*, Reston, VA, 1991.
21. National Council of Teachers of Mathematics, *Assessment Standards for School Mathematics*, Reston, VA, 1995.
22. National Research Council, *Everybody Counts*: *A Report to the Nation on the Future of Mathematics Education*, National Academy Press, Washington, DC, 1989.
23. K. H. Parshall and D. E. Rowe, *The Emergence of the American Mathematical Research Community, 1876–1900: J. J. Sylvester, Felix Klein, and E. H. Moore*, American Mathematical Society, Providence, RI, 1994.
24. J. Pelech and J. Parker, The graphing calculator and division of fractions, *Math. Teacher* 89 (1994), 304–305.
25. S. E. Sigurdson, The development of the idea of unified mathematics in the secondary school curriculum 1890–1930, Ph. D. dissertation, University of Wisconsin, Madison, 1962.
26. G. M. A. Stanic and J. Kilpatrick, Mathematics curriculum reform in the United States: A historical perspective, *Int. J. Educ. Res.* 17 (1992), 407–417.
27. L. A. Steen, *Calculus for a New Century*: *A Pump, Not a Filter* (MAA Notes No. 8), Mathematical Association of America, Washington, DC, 1987.
28. L. Steffe and T. Kieren, Radical constructivism and mathematics education, *J. Res. Math. Educ.* 25 (1994), 711–733.
29. E. von Glasersfeld, Steps in the construction of "others" and "reality": A study in self-regulation, in *Power, Autonomy, Utopia*, R. Trappl, ed., pp. 107–116, New York, Plenum, 1986.
30. E. von Glasersfeld, ed., *Radical Constructivism in Mathematics Education*, Dordrecht, Kluwer, 1991.
31. H. Wu, The role of open-ended problems in mathematics education, *J. Math. Behavior* 13 (1994), 115–128.
32. H. Wu, The role of Euclidean geometry in high school, *J. Math. Behavior* 15 (1996), 221–237.
33. H. Wu, The mathematics education reform: Why you should be concerned and what you can do, *Amer. Math. Monthly* 104 (1997), 946–954.

Department of Mathematics Education
105 Aderhold, University of Georgia
Athens, Georgia 30602-7124
jkilpat@coe.uga.edu

NOTES

Edited by **Jimmie D. Lawson and William Adkins**

A Shorter Proof of the Ramanujan Congruence Modulo 5

John L. Drost

A *partition* of a natural number n is a way of writing n as an unordered sum of natural numbers. The partition function $p(n)$ counts the number of partitions of n. For example, $p(4) = 5$, with $4, 3 + 1, 2 + 2, 2 + 1 + 1, 1 + 1 + 1 + 1$ being the five partitions of 4. The extraordinary Ramanujan at first noticed, then proved, some very interesting congruence properties of $p(n)$ modulo the primes 5, 7, and 11. In fact, he generalized them to congruences modulo powers of those primes. For the prime 5, Ramanujan's congruence is the following:

Theorem 1. *If m is a nonnegative integer, then* $p(5m + 4) \equiv 0 \pmod 5$.

The first few cases are $p(4) = 5, p(9) = 30, p(14) = 135$. The shortest proof of Theorem 1 in the literature can be found in [**1**; pp. 176–177] or [**2**; pp. 287–289]. They use both Euler's pentagonal number formula (for the inverse of the generating function of $p(n)$), and the following identity, due to Jacobi:

$$[(1-x)(1-x^2)(1-x^3)\cdots]^3$$
$$= 1 - 3x + 5x^3 - 7x^6 + \cdots + (-1)^n(2n+1)x^{n(n+1)/2} + \cdots \tag{1}$$

The congruence is then deduced by working in the power series ring modulo 5. A combinatorial proof of (1), using the Involution Principle, can be found in [**3**]; the classical proof uses the Jacobi triple product formula.

The proof given here is similar, but requires *only* Jacobi's identity and the binomial theorem. Along the way, a congruence result involving the square of the partition generating function is shown.

Proof of Theorem 1: The partition function has corresponding generating function given by $P(x) = [(1-x)(1-x^2)(1-x^3)\cdots]^{-1} = 1 + x + 2x^2 + 3x^3 + \cdots + p(n)x^n + \cdots$. Let $P_k(x)$ be $[P(x)]^k$ and let the x^n coefficient of $P_k(x)$ be $p_k(n)$. The first step in the proof is to show that $p_2(5m+i) \equiv 0 \pmod 5$ for all $m \geq 0$, and $i = 2, 3,$ or 4. To see this, note that the left side of (1) is $P_{-3}(x)$ from the product expansion of $P(x)$, and that $P_2(x) = P_{-3}(x)P_5(x) \equiv P_{-3}(x)P(x^5) \pmod 5$. In the congruence we are using the fact, which follows from the binomial theorem, that for all power series $f(x), [f(x)]^5 \equiv f(x^5) \pmod 5$. Expanding this last

product gives

$$P_2(x) \equiv (1 - 3x - 7x^6 + 9x^{10} - 11x^{15} + \cdots)$$
$$\times(1 + x^5 + 2x^{10} + 3x^{15} + \cdots) \pmod 5. \qquad (2)$$

The exponents of the first factor are all 0 or 1 mod 5, and the multiplication by the second factor does not change this. So the coefficients $p_2(n)$ mod 5 are nonzero only if $n \equiv 0$ or 1 (mod 5). In terms of power series, $P_2(x) \equiv r(x^5) + xs(x^5)$ (mod 5) for some power series $r(x)$, $s(x)$. To finish the proof, take the cube of both sides of the preceding congruence. This gives

$$P_6(x) \equiv r(x^5)^3 + 3xr(x^5)^2 s(x^5) + 3x^2 r(x^5) s(x^5)^2 + x^3 s(x^5)^3 \pmod 5.$$

Since $r(x^5)$ and $s(x^5)$ are power series in x^5, this implies that the x^{5m+4} coefficients of $P_6(x)$ are divisible by 5. Multiplying this last congruence by $P_{-5}(x)$, which is another power series in x^5 mod 5, does not alter this. ■

This proof was motivated by the classical one of the Ramanujan congruence mod 7, i.e., $p(7m + 5) \equiv 0 \pmod 7$. In that case one multiplies $P_{-3}(x)$ by $P_7(x)$. Modulo 7, one gets $P_4(x) \equiv t + xu + x^3 v$, where t, u, v are all power series in x^7. Squaring this gives $P_8(x)$ having coefficients of x^{7m+5} divisible by 7, which again won't change upon multiplying by $P_{-7}(x)$.

Alternate proofs can be obtained by replacing the cubing in the mod 5 case and the squaring in the mod 7 case by taking the square roots and fourth roots, respectively, of the formal power series. This gives the function $P(x)$ directly; however, there must now be an additional argument to show that the binomial coefficients generated (or the products of binomial coefficients in the mod 7 case) are 0 in the appropriate modulus.

REFERENCES

1. G. E. Andrews, *The Theory of Partitions*, Addison-Wesley, 1976.
2. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 5th ed., Oxford, 1979.
3. J. T. Joichi and D. Stanton, An involution for Jacobi's identity, *Discrete Math.* **73** (1989) 261–271.

Department of Mathematics
Marshall University
400 Hal Greer Blvd.
Huntington, WV 25755-2560
drost@marshall.edu

An Amusing Representation of $\dfrac{x}{\sin x}$

Scott Ahlgren, Lars English, and Ron Winters

In the course of their work [1] on the quantum mechanics of a simple harmonic oscillator, the latter two authors were confronted with a limit of products of partial

approximates of a continued fraction. With the aid of the symbolic manipulator Maple, they discovered that the quantity in question was nothing other than $x/\sin x$. In this Note we give a derivation of the resulting formula, which may be of interest to mathematicians and physicists alike.

When x is a non-zero real number and n is a positive integer with $n > |x|/\sqrt{2}$, define $\gamma(n,x) = 2 - x^2/n^2$. Set $C_0 = 0$ and $C_1 = 1/\gamma(n,x)$; define $C_p = (\gamma(n,x) - C_{p-1})^{-1}$ for $p = 2, 3, \dots$, so that

$$C_p = \underbrace{\cfrac{1}{\gamma(n,x) - \cfrac{1}{\cdots - \cfrac{1}{\gamma(n,x)}}}}_{p \text{ times}}$$

If x is not an integer multiple of π, we claim that

$$\lim_{n\to\infty} n \prod_{p=1}^{n-1} C_p = \frac{x}{\sin x}. \tag{1}$$

Our proof uses a potpourri of techniques from continued fractions, recurrence sequences, and calculus. To begin, set $A_0 = 0$ and $A_1 = 1$, and define $A_p = \gamma(n,x)A_{p-1} - A_{p-2}$ for $p = 2, 3, \dots$. An induction shows that $C_p = A_p/A_{p+1}$, so the product in (1) collapses, and the left side becomes simply $\lim_{n\to\infty} n/A_n$.

Let $\tau = \frac{1}{2}(\gamma(n,x) + \sqrt{(\gamma(n,x)^2 - 4)})$ and $\bar{\tau} = \frac{1}{2}(\gamma(n,x) - \sqrt{\gamma(n,x)^2 - 4})$ be the roots of the polynomial

$$X^2 - \gamma(n,x)X + 1. \tag{2}$$

Since $0 < \gamma(n,x) < 2$, we see that $\tau \neq \bar{\tau}$ and that $\bar{\tau}$ is the complex conjugate of τ. Using (2) and induction on p, we find that $A_p = (\tau^p - \bar{\tau}^p)/(\tau - \bar{\tau})$, and our limit becomes

$$\lim_{n\to\infty} \frac{n}{A_n} = \lim_{n\to\infty} \frac{n(\tau - \bar{\tau})}{\tau^n - \bar{\tau}^n}. \tag{3}$$

Notice that

$$\tau = 1 - \frac{x^2}{2n^2} + \frac{i|x|}{2n}\sqrt{4 - \frac{x^2}{n^2}},$$

from which we obtain $\lim_{n\to\infty} n(\tau - \bar{\tau}) = \lim_{n\to\infty} i|x|(4 - x^2/n^2)^{1/2} = 2i|x|$. Using l'Hôpital's rule, we find that $\lim_{n\to\infty} n \log \tau = i|x|$, whence $\lim_{n\to\infty} \tau^n = e^{i|x|}$. Therefore the limit in (3) is

$$\frac{2i|x|}{e^{i|x|} - e^{-i|x|}} = \frac{x}{\sin x}.$$

Notice that when $x = 0$ we have $A_n = n$; in this case the left side of (1) equals 1, as expected.

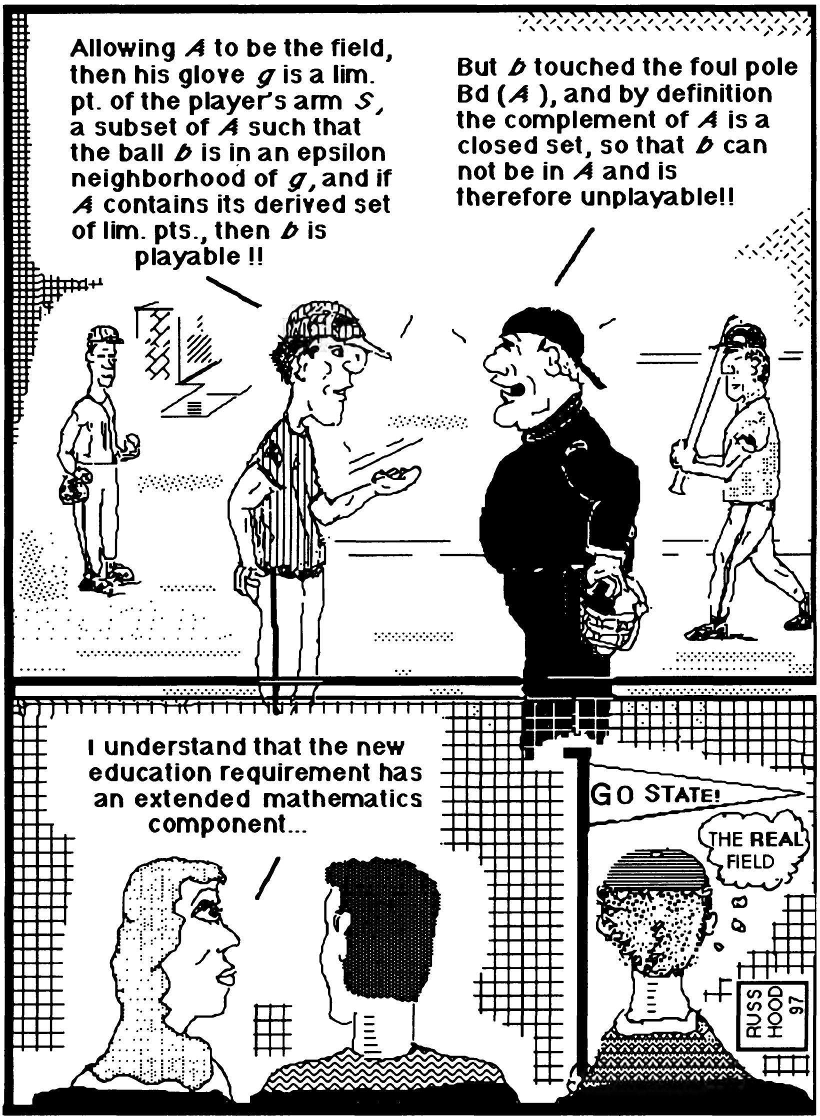
ACKNOWLEDGMENT. Ron Winters was supported in part by grant #P176B30079 from the Department of Education (FIPSE program) and a grant from the W.M. Keck Foundation.

REFERENCE

1. L. English and R. Winters, Continued fractions and the harmonic oscillator using Feynman's path integrals, *Am. J. Phys.* **65** (1997), 390–393

Ahlgren / Winters:
Department of Mathematics and Computer Science
/Department of Physics
Denison University
Granville, OH 43023
ahlgren@denison.edu / winters@denison.edu

English:
Department of Physics
Cornell University
Ithaca, NY 14850
quentin@tristan.tn.cornell.edu



Contributed by Russ Hood, Rio Linda, CA

PROBLEMS AND SOLUTIONS

Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**

with the collaboration of Paul T. Bateman, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Underwood Dudley, Michael A. Filaseta, Ira M. Gessel, Bart Goddard, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Robert Israel, Murray S. Klamkin, Daniel J. Kleitman, Fred Kochman, Frederick W. Luttmann, Frank B. Miles, Richard Pfiefer, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before May 31, 1998; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.*

PROBLEMS

10627. *Proposed by George E. Andrews, The Pennsylvania State University, University Park, PA.* The Rogers-Ramanujan (RR) partitions of an integer are those that have no repetitions and no consecutive integers as parts. The RR$'$ partitions are those RR partitions that have no 1's.
(a) For $n > 1$, prove that at least half of the RR partitions of n are RR$'$ partitions.
(b) Let $Q(n)$ denote the number of RR$'$ partitions of n into at least two parts whose two largest parts differ by at most 2 more than the number of parts. For example, $Q(12) = 3$ because, of the nine RR partitions of 12, six are RR$'$ partitions, and of these only three ($8+4$, $7+5$, and $6+4+2$) meet the stated condition. For $n > 1$, prove that $Q(n)$ equals the difference between twice the number of RR$'$ partitions of n and the number of RR partitions of n.

10628. *Proposed by George E. Andrews, The Pennsylvania State University, University Park, PA.* Let $p_{a,b}(n)$ denote the number of partitions of n that contain no parts of size a or b. For $n > 0$, prove that

$$\sum_{j\geq 1}(-1)^j p_{j,2j}\left(n - \frac{3j(j-1)}{2}\right) = 0.$$

For example, when $n = 9$ the assertion is $-p_{1,2}(9) + p_{2,4}(6) - p_{3,6}(0) = 0$, which is true because $p_{1,2}(9) = 4$ (the relevant partitions are 9, 6+3, 5+4, 3+3+3), $p_{2,4}(6) = 5$ (the relevant partitions are 6, 5+1, 3+3, 3+1+1+1, 1+1+1+1+1+1), and $p_{3,6}(0) = 1$ (the empty partition of 0 satisfies the condition).

10629. *Proposed by Frank Schmidt, Arlington, VA.* Let $p(n)$ denote the number of partitions of the integer n, and let $f(n)$ denote the number of partitions $\lambda_1 + \lambda_2 + \lambda_3 + \cdots$ satisfying $\lambda_1 > \lambda_2 > \lambda_3 > \cdots$ and $n = \lambda_1 + \lambda_3 + \lambda_5 + \cdots$. For example, $p(5)$ counts the 7 partitions $5, 4+1, 3+2, 3+1+1, 2+2+1, 2+1+1+1$, and $1+1+1+1+1$, and $f(5)$ counts the 7 partitions $5, 5+1, 5+2, 5+3, 5+4, 4+3+1$, and $4+2+1$. Prove that $p(n) = f(n)$ for every positive integer n.

10630. *Proposed by Richard Stong, Rice University, Houston, TX.* It is possible to show that $\csc(3\pi/29) - \csc(10\pi/29) = 1.999989433\ldots$. Prove that there are no integers j, k, n with n odd satisfying $\csc(j\pi/n) - \csc(k\pi/n) = 2$.

10631. *Proposed by Greg Huber, University of Chicago, Chicago, IL.* Given a triangle T, let the *intriangle* of T be the triangle whose vertices are the points where the circle inscribed in T touches T. Given a triangle T_0, form a sequence of triangles $T_0, T_1, T_2, \ldots$ in which each T_{n+1} is the intriangle of T_n. Let d_n be the distance between the incenters of T_n and T_{n+1}. Find $\lim_{n\to\infty} d_{n+1}/d_n$ when T_0 is not equilateral.

10632. *Proposed by William F. Trench, Trinity University, San Antonio, TX.* For given nonnegative integers m and n, evaluate

$$\sum_{k=0}^{m} \frac{(-1)^k}{n+k+1}\binom{m}{k}(1-y)^{n+k+1} + \sum_{k=0}^{n} \frac{(-1)^k}{m+k+1}\binom{n}{k}y^{m+k+1}.$$

10633. *Proposed by Kiran S. Kedlaya, Princeton University, Princeton, NJ.* Let S be a commuting family of n-by-n matrices over an arbitrary field. Suppose the matrices in M have a common eigenvector v, so that $Mv = \lambda_M v$ for all $M \in S$. Prove that the transposes of these matrices also have a common eigenvector with these eigenvalues, that is, a vector w satisfying $M^T w = \lambda_M w$ for all $M \in S$.

SOLUTIONS

A Partial Comparison Test for Divergence

10412 [1994, 911]. *Proposed by Donald A. Darling, Newport Beach, CA.* Find necessary and sufficient conditions on a nonincreasing sequence $a_1, a_2, \ldots$ of positive real numbers so that, if $b_1, b_2, \ldots$ is a nonincreasing sequence with $b_k \geq a_k$ for infinitely many k, then $\sum b_n = \infty$.

Solution by John H. Lindsey II, Fort Myers, FL. We need not assume the a's are nonincreasing. The sought condition is $\liminf na_n > 0$.

Suppose $\liminf na_n = 0$. Let $n_0 = 0$. Pick n_1 with $n_1 a_{n_1} < 1/2$, and then for each $k > 1$ pick $n_k > n_{k-1}$ with $n_k a_{n_k} < \min\left(n_{k-1}a_{n_{k-1}}, 2^{-k}\right)$. For $n_{k-1} < n \leq n_k$ define $b_n = a_{n_k}$. The b's are nonincreasing, and

$$\sum_{n=1}^{\infty} b_n = \sum_{k=1}^{\infty} \sum_{n=n_{k-1}+1}^{n_k} a_{n_k} = \sum_{k=1}^{\infty}(n_k - n_{k-1})a_{n_k} \leq \sum_{k=1}^{\infty} n_k a_{n_k} \leq \sum_{k=1}^{\infty} 2^{-k} = 1.$$

Suppose $\liminf na_n = 2c > 0$. Then there exists N so that $n > N$ implies $na_n > c$. If the b's satisfy the given conditions, we can define a sequence n_k with $n_0 = 0$, $n_1 > N$, $n_k > 2n_{k-1}$ for $k > 1$, and $b_{n_k} \geq a_{n_k}$. Then

$$\sum_{n=1}^{\infty} b_n \geq \sum_{k=1}^{\infty} \sum_{n=n_{k-1}+1}^{n_k} b_{n_k} \geq \sum_{k=1}^{\infty}(n_k - n_{k-1})a_{n_k} > \sum_{k=1}^{\infty} \frac{1}{2} n_k a_{n_k} \geq \sum_{k=1}^{\infty} \frac{1}{2}c = \infty.$$

Solved also by D. W. Bailey, R. Barbara (Lebanon), P. Budney, R. J. Chapman (U. K.), E. Hertz, R. Holzsager, M. Hudelson, N. Komanda, O. P. Lossers (The Netherlands), D. Marcus, A. Meir (Canada), H. Morris, G. Myerson (Australia), V. Novakov (Bulgaria), A. Pedersen (Denmark), C. G. Petalas & T. P. Vidalis (Greece), M. Reid, R. M. Robinson, K. Schilling, R. Stong, P. Szeptycki, A. A. Tarabay (Lebanon), D. R. Witte, A. N. 't Woord (The Netherlands), NSA Problems Group, and the proposer.

A Singular Inequality

10421 [1994, 1014]. *Proposed by Gigel Militaru, University of Bucharest, Bucharest, Romania.* Let n be an integer, $n \geq 3$, and let $z_1, \ldots, z_n$ and $t_1, \ldots, t_n$ be complex numbers. Prove that there exists an integer i with $1 \leq i \leq n$ such that $4\,|z_i t_i| \leq \sum_{j=1}^{n} |z_i t_j + z_j t_i|$.

Solution by O. P. Lossers, University of Technology, Eindhoven, The Netherlands. We prove the following generalization:

Proposition. *Let $a_1^{(l)}, \ldots, a_n^{(l)}$ and $b_1^{(l)}, \ldots, b_n^{(l)}$ $(l = 1, \ldots, k;\ k < n)$ be complex numbers. Then there exists an integer i with $1 \leq i \leq n$ such that*

$$2\left|a_i^{(1)} b_i^{(1)} + \cdots + a_i^{(k)} b_i^{(k)}\right| \leq \sum_{j=1}^{n} \left|a_i^{(1)} b_j^{(1)} + \cdots + a_i^{(k)} b_j^{(k)}\right|.$$

Proof. Let $M^{(l)}$ be the $n \times n$ matrix defined by $M_{ij}^{(l)} = (a_i^{(l)} b_j^{(l)})$ and define $M = \sum_{l=1}^{k} M^{(l)}$. The matrix $M^{(l)}$ has rank at most one, so $\operatorname{rank}(M) \leq k < n$. Being singular, M cannot be "diagonally dominant", so for some i we must have $|M_{ii}| \leq \sum_{j \neq i} |M_{ij}|$ or, removing the restriction on j, $2|M_{ii}| \leq \sum_{j=1}^{n} |M_{ij}|$, which gives the result. □

The desired inequality is obtained from the Proposition by taking $k = 2$, $a_i^{(1)} = b_i^{(2)} = z_i$, and $a_i^{(2)} = b_i^{(1)} = t_i$.

Editorial comment. For a proof of this property of diagonally dominant matrices, and more, see Roger A. Horn & Charles R. Johnson, *Matrix Analysis*, Cambridge, 1985, Theorem 6.1.10, p. 349.

Solved also by D. Beckwith, R. J. Chapman (U. K.), J. H. Lindsey II, R. Vermes (Canada), and the proposer.

Self-sorting in Tournaments

10447 [1995, 360]. *Proposed by Stephen C. Locke, Florida Atlantic University, Boca Raton, FL.* Consider a tournament in which every pair of teams played a match that one of the two won. Let L_0 be a listing of the teams in some order, and define successive L_i, $i = 1, 2, 3, \ldots$ by repeated application of the following operation: if a team in the list L_i lost to the team immediately following it in the list, call that pair of teams a *switchable pair*; the order of one switchable pair is then reversed to give L_{i+1}. Note that this may increase the number of switchable pairs.

Prove that any such sequence of operations leads, in a finite number of steps, to a list in which every team defeated the team immediately following it in the list, so there are no switchable pairs.

Composite solution by A. N. 't Woord, University of Technology, Eindhoven, The Netherlands, and Jerrold R. Griggs, University of South Carolina, Columbia, SC. Let a_i denote the number of pairs (A, B) where team A lost to team B, but team A precedes team B in list L_i. Then $a_{i+1} = a_i - 1$. Since all a_i are nonnegative integers, the process described must lead, after at most $\binom{n}{2}$ steps, to a list without switchable pairs. The maximum number of switches, $\binom{n}{2}$, is required if and only if we start with the reversed transitive tournament in which team j lost to team k for all j, k with $j < k$.

Solved also by D. Beckwith, K. L. Bernstein, R. J. Chapman (U. K.), B. Dawson, R. Ehrenborg & F. Fares (Canada), K. Foltz, S. M. Gagola Jr., F. Galvin, J. W. Grossman & R. S. Zeitman, C. Hillar, G. Isaak, N. Komanda, J. H. Lindsey II, J. B. Muskat (Israel), A. Nijenhuis, A. Pedersen (Denmark), P. J. Slater, J. H. Steelman, R. Stong, Anchorage Math Solutions Group, NSA Problems Group, Oklahoma State Problems Group, WMC Problems Group, and the proposer.

Polynomial Divisibility

10452 [1995, 463]. *Proposed by Seung-Jin Bang, Ajou University, Suwon, Korea.* Find all values of n, k, a, and b (n and k positive integers, $n > k$, a and b nonzero real numbers) for which the polynomial $x^n + ax + b$ is divisible by $x^k + ax + b$ in $\mathbb{R}[x]$.

Composite solution by Gerald A. Heuer, Concordia College, Moorhead, MN, and Roger B. Eggleton, Illinois State University, Normal, IL. The values are
(i) $k = 1$, $a = -1$, n arbitrary;
(ii) $k = 1$, $a \neq -1$, $b = -(a+1)$, n even;
(iii) $k = 1$, $a \neq -1$, $b = \pm(a+1)$, n odd;
(iv) $k = 2$, $b = 1$, $a = -2\cos(2r\pi/(n-2))$ for some integer r with $1 \le r \le \lfloor (n-3)/2 \rfloor$ and $r \neq (n-2)/4$, $n = 5$ or $n \ge 7$.

Let $P(x) = x^k + ax + b$. Note that $P(x)$ divides $x^n + ax + b$ if and only if it divides $x^n + ax + b - P(x) = x^n - x^k = (x^{n-k} - 1)x^k$. Since $b \neq 0$, this occurs when $P(x)$ divides $Q(x) = x^{n-k} - 1$.

We first consider the case $k = 1$. If $a = -1$, then $P(x) = b$, which divides $Q(x)$ for all $b \neq 0$. Otherwise, $P(x) = (a+1)x + b = \big(a+1\big)\big(x + b/(a+1)\big)$, so $P(x)$ divides $Q(x)$ if and only if $\zeta = -b/(a+1)$ satisfies $\zeta^{n-1} = 1$. Since ζ is real, this occurs when $\zeta = 1$ or when $\zeta = -1$ and $n - 1$ is even. This explains cases **(i)**–**(iii)**.

Now suppose $k > 1$ and $P(x)$ divides $Q(x)$. Since $P(x)$ is real, its nonreal roots come in complex conjugate pairs, so its irreducible factors over $\mathbb{R}$ can be only $x - 1$, $x + 1$, or

$$\big(x - e^{2r\pi i/(n-k)}\big)\big(x - e^{-2r\pi i/(n-k)}\big) = x^2 - 2x\cos\tfrac{2r\pi}{n-k} + 1$$

for some r with $0 < r < (n-k)/2$. Note that $x - 1$ is a negative-reciprocal polynomial and all other possible irreducible factors are reciprocal polynomials, so $P(x)$ is either a reciprocal polynomial or a negative-reciprocal polynomial. It follows that

$$x^k + ax + b = P(x) = \pm x^k P(1/x) = \pm(bx^k + ax^{k-1} + 1).$$

Since $a \neq 0$, comparing coefficients of x shows that $k = 2$ and the plus sign occurs. Now, comparing coefficients of x^k shows that $b = 1$. Since $Q(x)$ has no multiple roots, $P(x)$ cannot be $(x-1)^2$ or $(x+1)^2$, so $P(x)$ must be $x^2 - 2x\cos(2r\pi/(n-k)) + 1$ for some r satisfying $1 \le r \le \lfloor (n-3)/2 \rfloor$. Note that $r = (n-2)/4$ gives $a = 0$, which is forbidden. All other values of r give solutions. There are no allowable values of r unless $n = 5$ or $n \ge 7$.

Solved also by M.-Th. Antoine (France), R. Barbara (Lebanon), M. Benedicty, D. Callan, R. J. Chapman (U. K.), F. J. Flanigan, Z. Franco, S. M. Gagola Jr., A. Gunawardena, N. Komanda, J. H. Lindsey II, O. P. Lossers (The Netherlands), A. Pedersen (Denmark), Y. Wang, A. N. 't Woord (The Netherlands), Anchorage Math Solutions Group, NSA Problems Group, and the proposer.

More On Stirling's Approximation

10521 [1996, 347]. *Proposed by D. M. Bloom & G. W. Booth, Brooklyn College, CUNY, Brooklyn, NY.* Let $S_n = (2\pi n)^{1/2}(n/e)^n$ and $T_n = n!/S_n$.
(a) Prove that $T_n - 1 = 1/(12n - a_n)$, where $0 < a_n < \frac{1}{2}$ for all positive integers n.
(b) Prove that the sequence $\langle a \rangle$ is monotonically increasing.
(c)* If $b_n = n(1/2 - a_n)$ for all $n \in \mathbb{N}$, is the sequence $\langle b \rangle$ monotonically increasing?

Solution by Richard Stong, Rice University, Houston, TX. The sequence b_n is increasing. This, as well as parts **(a)** and **(b)**, follows from the asymptotics of the gamma function carried through with error bounds. For these error bounds, we use the following convention: Any appearance of the letter θ or τ in a formula means that the formula holds if θ is replaced by some number (possibly different for each appearance) in $(0, 1)$ and τ is replaced by some number in $(-1, 1)$.

Note that $\log(T_n/T_{n+1}) = (n+1/2)\log(1+1/n) - 1$. Stirling's formula ensures that $T_n \to 1$, so we have $T_n = \exp(R_n)$, where $R_n = \sum_{k=n}^{\infty}\left((k+1/2)\log(1+1/k) - 1\right)$. Expanding in powers of $1/k$ gives

$$\begin{aligned}
&\frac{1}{12}\left(\frac{1}{k} - \frac{1}{k+1}\right) - \frac{1}{360}\left(\frac{1}{k^3} - \frac{1}{(k+1)^3}\right) \\
&\qquad + \frac{1}{1260}\left(\frac{1}{k^5} - \frac{1}{(k+1)^5}\right) - \left(\left(k+\frac{1}{2}\right)\log\left(1+\frac{1}{k}\right) - 1\right) \\
&\qquad = \sum_{r=8}^{\infty}(-1)^r\left(\frac{1}{1260}\binom{r-1}{4} - \frac{1}{360}\binom{r-1}{2} + \frac{1}{12} - \frac{r-1}{2r(r+1)}\right)k^{-r} \\
&\qquad = \frac{1}{240k^8} - \frac{1}{60k^9} + \cdots. \qquad (1)
\end{aligned}$$

(The coefficients have been chosen to cancel the terms with $r \le 7$.) The summands in (1) alternate in sign, converge to zero, and are decreasing in magnitude for $k \ge 4$. Therefore,

$$\begin{aligned}
\left(k+\frac{1}{2}\right)\log\left(1+\frac{1}{k}\right) - 1 &= \frac{1}{12}\left(\frac{1}{k} - \frac{1}{k+1}\right) - \frac{1}{360}\left(\frac{1}{k^3} - \frac{1}{(k+1)^3}\right) \\
&\qquad + \frac{1}{1260}\left(\frac{1}{k^5} - \frac{1}{(k+1)^5}\right) - \frac{\theta}{240k^8}.
\end{aligned}$$

Summing from n to ∞ gives, for $n \ge 4$,

$$R_n = \frac{1}{12n} - \frac{1}{360n^3} + \frac{1}{1260n^5} - \frac{11\theta}{6720n^7}, \qquad (2)$$

where we obtain the last term from the upper bound

$$\sum_{k=n}^{\infty}\frac{1}{240k^8} \le \frac{1}{240n^8} + \int_n^{\infty}\frac{1}{240x^8}\,dx = \frac{1}{240n^8} + \frac{1}{1680n^7} \le \frac{11}{6720n^7}.$$

The terms in (2) are decreasing in magnitude for $n \ge 4$, so we can truncate at any step to get an inequality involving R_n. For $n \ge 4$ we get $1/(13n) \le R_n \le 1/(12n)$, and using this to bound derivatives gives

$$R_n^3 = \left(\frac{1}{12n} - \frac{\theta}{360n^3}\right)^3 = \frac{1}{1728n^3} - \frac{\theta}{17280n^5},$$

$$R_n^5 = \left(\frac{\theta}{12n}\right)^5 = \frac{\theta}{248832n^5},$$

and

$$\begin{aligned}
\frac{1}{R_n} &= \frac{12n}{1 - 1/(30n^2) + 1/(105n^4)} + \frac{1859\theta}{6720n^5} \\
&= 12n\left(1 + \frac{1}{30n^2} - \frac{1}{105n^4} + \left(\frac{1}{30n^2} - \frac{1}{105n^4}\right)^2 + \theta\left(\frac{1}{30n^2} - \frac{1}{105n^4}\right)^3\right) + \frac{1859\theta}{6720n^5} \\
&= 12n + \frac{2}{5n} - \frac{53}{525n^3} - \frac{4}{525n^5} + \frac{4}{3675n^7} + \frac{\theta}{2250n^5} + \frac{1859\theta}{6720n^5} \\
&= 12n + \frac{2}{5n} - \frac{53}{525n^3} + \frac{0.269\tau}{n^5}.
\end{aligned}$$

From E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Cambridge University Press, Cambridge, 1927, Equations 7.2 and 13.151, we have

$$\begin{aligned}
\frac{1}{T_n - 1} = \frac{1}{e^{R_n} - 1} &= \frac{1}{R_n} - \frac{1}{2} + \frac{R_n}{12} - \frac{R_n^3}{720} + \frac{R_n^5}{30240} - \cdots \\
&= \frac{1}{R_n} - \frac{1}{2} + \sum_{m=1}^{\infty}\frac{(-1)^{m-1}\zeta(2m)}{2^{2m-1}\pi^{2m}}R_n^{2m-1}. \qquad (3)
\end{aligned}$$

The terms in (3) alternate in sign, tend to zero, and (since the zeta function is decreasing and $R_n \le 2$) are decreasing. Therefore we can truncate (3) to get inequalities. Hence we see

$$\frac{1}{T_n - 1} = \frac{1}{R_n} - \frac{1}{2} + \frac{R_n}{12} - \frac{R_n^3}{720} + \frac{R_n^5 \theta}{30240}.$$

Plugging in our formulas for powers of R_n gives

$$\frac{1}{T_n - 1} = 12n - \frac{1}{2} + \frac{293}{720n} - \frac{4406147}{43545600n^3} + \frac{0.27\tau}{n^5}. \tag{4}$$

To prove parts (**a**) and (**b**), note that (4) translates into

$$a_n = \frac{1}{2} - \frac{293}{720n} + \frac{0.12\theta}{n^3}$$

for $n \ge 4$, so

$$a_{n+1} - a_n = \frac{293}{720n(n+1)} + \frac{0.12\tau}{n^3} > 0.$$

One also calculates $a_1 = 0.1569\ldots$, $a_2 = 0.3073\ldots$, $a_3 = 0.3678\ldots$, and $a_4 = 0.3997\ldots$. Therefore, we see that the a_n are increasing for all n and stay between 0 and their upper limit $1/2$. For part (**c**), (4) gives (for $n \ge 4$)

$$b_n = \frac{293}{720} - \frac{4406147}{43545600n^2} + \frac{0.27\tau}{n^4},$$

so

$$b_{n+1} - b_n = \frac{4406147(2n+1)}{43545600n^2(n+1)^2} + \frac{0.54\tau}{n^4} > 0.$$

Moreover, $b_1 = 0.3430\ldots$, $b_2 = 0.3853\ldots$, $b_3 = 0.3965\ldots$, and $b_4 = 0.4009\ldots$, so the b_n are increasing for all n.

Solved also by J. Anglesio (France), P. Bracken, J. S. Frame, J. H. Lindsey II, R. Richberg (Germany), G. Rzadkowski (Poland), Z. Sasvári (Germany), GCHQ Problems Group (U. K.), and the proposers (parts a and b).

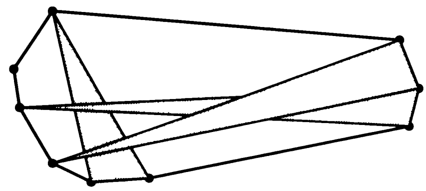
Balanced and Unbalanced Polygons

10526 [1996, 427]. *Proposed by Harry Tamvakis, University of Chicago, Chicago, IL.* Let $P = A_1A_2\ldots A_n$ be a convex polygon. For any point M in the interior, let B_i be the point where A_iM intersects the perimeter. We say that P is *balanced* if for some such M the points $B_1, B_2, \ldots, B_n$ are interior to distinct sides of P. Prove or disprove:
(**a**) If n is even, then P is not balanced.
(**b**) If n is odd, then P is balanced.

Solution by Mark Bowron and Stanley Rabinowitz, MathPro Press, Westford, MA. We prove (**a**) and disprove (**b**).
(**a**) Suppose n is even. Pick any two opposite vertices of P, say A_1 and A_m where $m = n/2 + 1$. The diagonal joining them divides P into two halves $H_1 = A_1A_2\ldots A_m$ and $H_2 = A_mA_{m+1}\ldots A_nA_1$. Suppose M is a balance point (so that $B_1, B_2, \ldots, B_n$ are interior to distinct sides of P). Note that M cannot lie on A_1A_m since $B_1 \ne A_m$. Thus M is interior to either H_1 or H_2. If M is interior to H_2, then the m points $B_1, B_2, \ldots, B_m$ must lie on the $m-1$ sides of $P \cap H_2$ since P is convex. By the pigeonhole principle, some side of P therefore contains more than one B_i. This is a contradiction. A similar contradiction arises if M is interior to H_1. Therefore P is not balanced.
(**b**) We give a counterexample (with $n = 9$) to show that P need not be balanced. Let a_i denote the side of the polygon opposite vertex A_i. If M is a balance point, then B_i must lie on a_i. Otherwise A_iB_i would divide P into two halves, with one half having more vertices (from which balance rays must emanate) than the other half has edges (to which balance

rays must terminate). Thus for all i, the point M must lie inside the triangle with vertex A_i and opposite side a_i.

If the polygon in the figure contains a balance point, that point must lie in each of the three shaded triangles. But these triangles have no point in common. Hence the polygon is not balanced.



An unbalanced 9-gon.

Editorial comment. All convex n-gons for which n is odd and $n \leq 7$ are balanced. For triangles, any interior point is a balance point. For convex pentagons, any point within the interior pentagon formed by the diagonals is a balance point. Helly's Theorem can be applied to show that all convex heptagons are balanced.

Solved also by G. L. Body (U. K.), R. J. Chapman (U. K.), D. A. Darling, J. E. Dawson (Australia), H. Guggenheimer (part a only), E. Gutkin (part a only), D. C. Kay, J. H. Lindsey II, M. D. Meyerson, C. Popescu (Belgium), H. Sedinger, GCHQ Problems Group (U. K.), and the proposer.

Is It An Integer?

10527 [1996, 427]. *Proposed by Vicenţiu Paşol, Craiova, Romania.* For positive integers m and n, let

$$N = 2\int_0^{\pi/2} \sin^{2n-1/2}\theta \cdot \cos^{2m+1/2}\theta \, d\theta.$$

Prove that $\sin\left(\left(2^{3(m+n)-2}\right)\cdot N \cdot \sqrt{2}\right) = 0$.

Composite solution by Allen Stenger, Tustin, CA, and the late J. Sutherland Frame, East Lansing, MI. Write $A = 2^{3(m+n)-2} \cdot N \cdot \sqrt{2}/\pi$. The stated result is a roundabout way of saying that A is an integer. We exhibit A as a quotient of integers and show that each prime divides the numerator in this quotient to at least as high a power as it does the denominator.

We use two Γ-function formulas in E. T. Whittaker and G. N. Watson, *A Course in Modern Analysis*, Cambridge, 1927, sections 12.42 and 12.14. The first is $N = \Gamma(m+3/4)$ $\Gamma(n+1/4)/\Gamma(m+n+1)$. The second is the functional equation $\Gamma(z)\Gamma(1-z) = \pi/\sin\pi z$ in the particular case $z = 1/4$, which yields $\Gamma(1/4)\Gamma(3/4) = \pi\sqrt{2}$. This gives

$$\begin{aligned} A &= 2^{3(m+n)-2}\,\frac{\dfrac{4m-1}{4}\dfrac{4m-5}{4}\cdots\dfrac{3}{4}\Gamma\left(\dfrac{3}{4}\right)\dfrac{4n-3}{4}\dfrac{4n-7}{4}\cdots\dfrac{1}{4}\Gamma\left(\dfrac{1}{4}\right)}{(m+n)!}\,\frac{\sqrt{2}}{\pi} \\ &= 2^{m+n-2}\,\frac{(4m-1)(4m-5)\cdots 3\cdot(4n-3)(4n-7)\cdots 1}{(m+n)!}\pi\sqrt{2}\frac{\sqrt{2}}{\pi} \\ &= 2^{m+n-1}\,\frac{(4m-1)(4m-5)\cdots 3\cdot(4n-3)(4n-7)\cdots 1}{(m+n)!} \end{aligned}$$

The exponent of the highest power of p dividing $n!$ is $\sum_{k\geq 1}\lfloor n/p^k\rfloor$. For $p = 2$ the exponent of the power of 2 in the denominator of A is

$$\sum_{k\geq 1}\left\lfloor\frac{m+n}{2^k}\right\rfloor < \sum_{k\geq 1}\frac{m+n}{2^k} = m+n,$$

hence is not greater than $m+n-1$, the exponent of the power of 2 in the numerator.

Now we consider the odd primes. Writing $A = A(m,n)$ as

$$2^{m+n-1}(-1)^m \frac{(1-4m)(5-4m)\cdots(-3)\cdot 1\cdot 5\cdots(4n-7)(4n-3)}{(m+n)!},$$

we see that the numerator is a product of $m+n$ consecutive terms in an arithmetic progression of difference 4. For any (odd) p^k, exactly one in every consecutive p^k terms of this product is divisible by p^k, so there are either $\lfloor (m+n)/p^k \rfloor$ or $\lceil (m+n)/p^k \rceil$ multiples of p^k in the product. Thus the power to which p divides the numerator is at least as large as the power to which it divides the denominator.

Editorial comment. Another way to proceed after obtaining the rational formula for A is to establish that $A(m,m)$ and $A(m, m+1)$ are integers and then invoke a recurrence relation. Both Jean Anglesio and Calin Popescu converted the original integral for N to the form

$$N = N(m,n) = 4\int_0^\infty \frac{t^{4n}}{(1+t^4)^{m+n+1}}\,dt \qquad (*)$$

with the change of variable $t = \tan\theta$. Anglesio then evaluated $N(m,0)$ and $N(m,1)$ by differentiating $N_m(x) := 4\int_0^\infty dt/(x^4+t^4)^{m+1}$ with respect to x repeatedly, and obtained the product for A by a recurrence relation. Popescu converted the integral in $(*)$ to a sum of multiples of integrals of the form $\int_0^\infty (t^4+1)^{-k}\,dt$ and evaluated these by contour integration and residues.

Solved also by J. Anglesio (France), G. Bach (Germany), D. Beckwith, D. Borwein & G. Sinnamon (Canada), R. J. Chapman (U. K.), J. E. Dawson (Australia), R. A. Groeneveld, W. Janous (Austria), L. E. Mattics, C. Popescu (Belgium), R. Richberg (Germany), V. Schindler (Germany), T. V. Trif (Romania), GCHQ Problems Group (U. K.), NSA Problems Group, and the proposer.

The Spiral of Cornu Has No Self-Intersections

10530 [1996, 509]. *Proposed by Daniel Goffinet, Saint Étienne, France.* The Cornu spiral in the complex plane is defined by the parameterization

$$t \mapsto z(t) = \int_0^t e^{i\pi u^2/2}\,du.$$

The eye sees no self-intersections. Is this a correct observation?

Solution by Joel Zeitlin, California State University, Northridge, CA. Note that $z(t)$ is above the real axis for $t > 0$ and below for $t < 0$, so it suffices to consider only $t > 0$. Also, $z(t)$ is a unit speed curve with strictly increasing curvature $k(t) = \pi t$. This implies that the osculating circles are strictly nested; see J. J. Stoker, *Differential Geometry*, Wiley-Interscience, New York, 1969, p. 31, or J. Zeitlin, Nesting behavior of osculating circles and the Fresnel integrals, *Math. Mag.* 54 (1981) 76–78. Since each point lies on its own osculating circle, it cannot coincide with any other point on the curve.

Editorial comment. Some solvers showed that the distance to the limit point is strictly decreasing. Others showed that the integral from a to b is non-zero for any $a < b$.

Solved also by K. Andersen, N. Blachman, R. J. Chapman (U. K.), D. Constales (Belgium), D. A. Darling, J. E. Dawson (Australia), E. Heil (Germany), J. H. Lindsey II, M. Omarjee (France), P. Walker (Oman), GCHQ Problems Group (U. K.), Wilmer Alabama Mathematics Club, and the proposer.

Choosing Random Numbers Until the Sum Meets a Threshold

10531 [1996, 510]. *Proposed by Emeric Deutsch, Polytechnic University, Brooklyn, NY, and Ieda Rodrigues, Cleveland State University, Cleveland, OH.* Let $x > 0$. Show that

$$\sum_{q=0}^{\lfloor x \rfloor} \frac{(-1)^q (x-q)^q e^{x-q}}{q!} < 2x+1.$$

Solution by Lise Jensen, Northeastern Illinois University, Chicago, IL. Let $f(x)$ denote the function on the left side of the inequality. Then **(i)** $f(x) = e^x$ if $0 < x < 1$ and $f(x) = e^x - (x-1)e^{x-1}$ if $1 \le x < 2$, **(ii)** $f(x)$ is continuous for all x, and **(iii)** for all $x > 1$,

$$\int_{x-1}^{x} f(t)\,dt = f(x) - 1\,. \tag{1}$$

Statement **(i)** follows directly from the expression for f and **(ii)** follows from observing that if x is not an integer, then the summand is continuous and if x is an integer and $q = x$, then the summand is 0. Statement **(iii)** can be seen formally (and proved rigorously) by observing that the Laplace transform of the function given by (i) and (iii) is $1/(s-1) + e^{-s} = \sum_{q=0}^{\infty}(-1)^q(s-1)^{-(q+1)}e^{-qs}$ and that $(s-1)^{-(q+1)}e^{-qs}$ is the Laplace transform of $\theta_q(x)(x-q)^q e^{x-q}/q!$ where $\theta_q(x) = 1$ if $x \ge q$ and $= 0$ otherwise. It can also be proved directly by integrating by parts, shifting indices, and using induction.

Consider the linear functional equation

$$F(x) = 1 + \int_{x-1}^{x} F(t)\,dt\,. \tag{2}$$

For every g continuous in $(0, 1]$, (2) has a unique solution F_g such that $F_g = g$ in the interval $(0, 1]$. To prove this, note that if F is known for $0 < x \le N+1$, then for $N < x \le N+1$ we have $F(x) = \phi(x) + \int_N^x F$, where ϕ is a known function; this can be solved for F in terms of ϕ. Also, note that F_g is positive if g is positive and that $F_g > F_h$ if $g(x) > h(x) > 0$ for all x in $(0, 1]$. The function $1 + 2x$ satisfies Equation (2), so $F_{1+2x}(x) = 1 + 2x$ and $F_{e^x}(x) = f(x)$. Now $e^x < 1 + 2x$ for $0 < x \le 1$, since $e^x - 1 - 2x$ is 0 at 0, is negative at 1, and has its only minimum at $\log 2$. It follows that $f(x) < 1 + 2x$ for all $x > 0$.

Editorial comment. The proposers explain the problem in the following way. Suppose one keeps selecting (independently) random numbers uniformly distributed in the interval $[0, 1]$ until the total exceeds x. The number of selections is a random variable Z_x having possible values $1, 2, 3, \ldots$. The expected value of Z_x turns out to be the sum in the problem.

Solved also by D. A. Darling, L. E. Mattics, GCHQ Problems Group (U. K.), and the proposers.

The Iterated Sine Sequence

10535 [1996, 510]. *Proposed by Vladimir Janković and Jovan Vukmirović, Belgrade, Yugoslavia.* Given s_0 with $0 < s_0 < \pi/2$, use $s_{n+1} = \sin s_n$ to define the sequence $\langle s \rangle$. Show that $n^2 s_n^2 - 3n + (9/5)\ln n$ is convergent.

Solution by Robin J. Chapman, University of Exeter, Exeter, UK. Since $0 < \sin x < x$ if $0 < x < \pi/2$, it follows that $\langle s \rangle$ is a decreasing and bounded sequence. Its limit s satisfies $\sin s = s$, and so $s = 0$. From the Maclaurin series for the sine function,

$$s_{n+1} = s_n\left(1 - \frac{s_n^2}{6} + \frac{s_n^4}{120} + O(s_n^6)\right).$$

Setting $u_n = 1/s_n^2$ gives

$$u_{n+1} = u_n\left(1 + \frac{1}{3u_n} + \frac{1}{15u_n^2} + O(u_n^{-3})\right). \tag{$*$}$$

Since $u_n \to \infty$ as $n \to \infty$ it follows from $(*)$ that for large enough n we have $u_{n+1} - u_n > 1/4$, and so $u_n > n/4 - A$ for some constant A. Hence $u_{n+1} - u_n = 1/3 + O(n^{-1})$. Summing this gives $u_n = n/3 + O(\log n)$, and so $1/u_n = 3/n + O\left((\log n)/n^2\right)$. Inserting this in $(*)$ gives

$$u_{n+1} - u_n = \frac{1}{3} + \frac{1}{5n} + O\left(\frac{\log n}{n^2}\right).$$

Since $\sum_{n>1} n^{-2}\log n$ is convergent and $\sum_{j=1}^{n} j^{-1} = \log n + \gamma + o(1)$ (where γ is Euler's constant), we have

$$u_n = \frac{n}{3} + \frac{\log n}{5} + K + o(1)$$

for some constant K. Hence

$$n^2 s_n^2 = \frac{n^2}{u_n} = 3n - \frac{9\log n}{5} - 9K + o(1),$$

and so $n^2 s_n^2 - 3n + \frac{9}{5}\log n \to -9K$ as $n \to \infty$.

Editorial comment. Several solvers noted that the iterated sine sequence of this problem is studied in N. G. de Bruijn, *Asymptotic Methods in Analysis*, Dover, 1981. On p. 159, de Bruijn proves the stronger statement

$$s_n = \sqrt{\frac{3}{n}}\left(1 - \frac{3}{10}\frac{\log n}{n} - \frac{C}{2n} + \frac{\alpha \log^2 n + \beta \log n + \gamma}{n^2} + O\left(\frac{\log^3 n}{n^3}\right)\right),$$

where C depends on s_0 and where α, β, and γ are explicit polynomials in C.

Solved also by U. Abel (Germany), S. Amghibech (France), J. Anglesio (France), D. Constales (Belgium), D. A. Darling, D. Doster, L. Jensen, K.-W. Lau (Hong Kong), J. H. Lindsey II, L. E. Mattics, M. Omarjee (France), A. Stenger, P. Walker, GCHQ Problems Group (U. K.), and the proposers.

A Pair of Exponential Equations

10537 [1996, 598]. *Proposed by Jonathan Aronson, Carnegie Mellon University, Pittsburgh, PA.* (**a**) For which positive real numbers c do there exist real numbers A and B with $B = ce^{-A}$, $A = ce^{-B}$, and $A \neq B$? (**b**) Show that $AB < 1$ when such A and B exist.

Composite solution I by Gilbert N. Lewis, Michigan Technological University, Houghton, MI, and Erik Doeff, Montana State University, Bozeman, MT. We must have $A, B > 0$, and we may assume $A > B > 0$. A solution to $B = ce^{-A}$ and $A = ce^{-B}$ is equivalent to a solution to $Ae^{-A} = Be^{-B}$ and $c = Be^{A} = Ae^{B}$. The function $f(x) = xe^{-x}$ is increasing on $(0, 1)$ and decreasing on $(1, \infty)$. Hence $A > 1 > B$, and B is implicitly defined as a function of A. Integrating $(x-1)^2/x^2 = 1 - 2/x + 1/x^2 > 0$ on $(1, \infty)$ shows that $x - 2\ln x - 1/x > 0$ on $(1, \infty)$. Exponentiating and rearranging gives $e^{-1/x}/x > xe^{-x}$. Hence $f(1/A) > f(A) = f(B)$. Since f is increasing on $(0, 1)$, this implies $1/A > B$ or $AB < 1$.

Now $\ln A - A = \ln B - B$, so differentiating implicitly gives

$$\frac{dB}{dA} = \frac{(1-A)B}{A(1-B)}$$

and

$$\frac{dc}{dA} = \left(B + \frac{dB}{dA}\right)e^{A} = \frac{1-AB}{A(1-B)}Be^{A} > 0.$$

Hence c is an increasing function of A for $A \in (1, \infty)$. As $A \to 1^+$, $B \to 1^-$ and hence $c \to e$. As $A \to \infty$, $c = Ae^{B} > A$ also tends to ∞. Thus c attains exactly the values in the range (e, ∞).

Solution II by Harry Sedinger, St. Bonaventure University, St. Bonaventure, NY. Let $g(x) = ce^{-x}$. The existence of A and B is equivalent to having two fixed points of $f = g \circ g$ other than the unique fixed point P of g. Other than P, the fixed points of $g \circ g$ come in pairs A and $g(A)$. Note that A and $g(A)$ occur on opposite sides of P, hence $A \neq g(A)$. Now $f'(x) = c^2 e^{-(x+g(x))} > 0$ and $f''(x) = c^2 e^{-(x+g(x))}(-1 + ce^{-x})$. Therefore f is increasing, concave upward for $x < \ln c$, and concave downward for $x > \ln c$. Also note

that $f(0) > 0$ and $f(x) \to c$ as $x \to \infty$. There are two cases: (**i**) P is the unique fixed point of f and $f'(P) \leq 1$, and (**ii**) there are three fixed points of f, $A < P < B$, $f'(A) < 1$, $f'(P) > 1$, and $f'(B) < 1$.

Note that $f'(P) = c^2e^{-2P} = P^2$, so (**i**) occurs when $P \leq 1$, or equivalently when $c \leq e$, and (**ii**) occurs when $P > 1$, or equivalently when $c > e$. This solves (**a**): a solution exists if and only if $c > e$. To see (**b**), note that in case (**ii**), $1 > f'(A) = c^2e^{-(A+g(A))} = ce^{-A}ce^{-B} = BA$.

Editorial comment. Frank J. Flanigan noted that part (**a**) is equivalent to problem 1313 in *Math. Mag.* [1989, 58; 1990, 59], which asks for pairs of points (x, y) on the exponential $y = a^x$ that are symmetric about the line $y = x$. If A, B, and c satisfy the equations of the problem, then the points $(A/c, B/c)$ and $(B/c, A/c)$ lie on $y = (e^{-c})^x$, and conversely. Can A. Minh noted that part (**b**) appeared as problem 27.8 in *Math. Spectrum* 28 No. 1 (1995/6). Several solvers noted that one can avoid implicit differentiation in solution I by setting $r = B/A$ and solving for A, B, and c as functions of r.

Solved also by P. Alsholm (Denmark), S. Amghibech (France), K. .F. Andersen (Canada), J. Anglesio (France), R. Barbara (Lebanon), M. Brozinsky, P. Budney, D. Callan, R. J. Chapman (U. K.), G. G. Chappell, D. Constales (Belgium), D. A. Darling, J. E. Dawson (Australia), F. J. Flannigan, K. Ford, J. Gutiérrez (Spain), J. L. Hartman, G. Isaacs, R. A. Kopas, O. Kraft & M. Schaefer (Germany), G. Lafferriere, J. H. Lindsey II, R. Manning, A. Nijenhuis, K. Schilling, H.-J. Seiffert (Germany), M. Shemesh (Israel), W. R. Smythe, R. Weinstock, L. Widmer, Anchorage Math Solutions Group, Con Amore Problems Group (Denmark), GCHQ Problems Group (U. K.), NCCU Problems Group, NSA Problems Group, Wilmer Alabama Mathematics Club, and the proposer.

REVIVALS

A Special Sequence of Algebraic Integers

E 3461 [1991, 755; 1997, 171]. *Proposed by David Callan, University of Wisconsin, Madison, WI.* Suppose r is a rational number but not an integer. It is known that $\tan(r\pi/2)$ is an algebraic number; see Ivan Niven, *Irrational Numbers*, Carus Mathematical Monographs No. 11, pp. 37-41. Find the smallest positive integer k_r such that $k_r \tan(r\pi/2)$ is an algebraic integer.

Editorial comment. Unfortunately, there were two minor errors in the published solution.

First, the opening sentence of the solution was inaccurately worded. It should have read: "If the denominator of $r/2$ in lowest terms is twice a power of an odd prime p, then $k_r = p$; otherwise $k_r = 1$." In other words, for $k_r \neq 1$, one needs not only that the denominator of r is a power of an odd prime but also that the numerator of r is odd.

Second, it was asserted that, for $n > 2$, irreducibility of the cyclotomic polynomial $\Phi_n(z)$ over $\mathbb{Q}$ implies irreducibility over $\mathbb{Q}$ of the polynomial

$$P_n(t) = (1 - it)^{\phi(n)}\Phi_n\left(\frac{1 + it}{1 - it}\right).$$

The example $\Phi_4(z) = z^2 + 1$, $P_4(t) = 2 - 2t^2$ shows that this argument is specious. However, irreducibility of $P_n(t)$ is really needed only when n is twice a power of an odd prime. If p is prime and α is a positive integer, then $\Phi_{2p^\alpha}(z) = (1 + z^{p^\alpha})/(1 + z^{p^{\alpha-1}})$, so

$$P_{2p^\alpha}(t) = \frac{(1 + it)^{p^\alpha} + (1 - it)^{p^\alpha}}{(1 + it)^{p^{\alpha-1}} + (1 - it)^{p^{\alpha-1}}} = \frac{2\big(1 + pA(t)\big)}{2\big(1 + pB(t)\big)} = 1 + pC(t),$$

where $A(t)$, $B(t)$, and $C(t)$ are polynomials with integer coefficients, and the leading coefficient of $C(t)$ is $(-1)^{(p-1)/2}$. Irreducibility of $P_{2p^\alpha}(t)$ then follows from the Eisenstein Irreducibility Theorem; see Harry Pollard and Harold G. Diamond, *The Theory of Algebraic Numbers*, Carus Mathematical Monographs, No. 9, pp. 30–35.

REVIEWS

Edited by **Underwood Dudley**
Mathematics Department, De Pauw University, Greencastle, IN 46135

Conceptual Mathematics: A First Introduction to Categories. By F. William Lawvere & Steven Schanuel with the cooperation of Emilio Faro, Fatima Fenaroli, Danilo Lawvere and the students of Mathematics 108 of the University of Buffalo. Cambridge University Press, 1997, 358, $80.00 (hard bound), $25.95 (paper bound).

Reviewed by **Saunders Mac Lane**

The amazing science of mathematics involves a mixture of calculations and concepts. When we teach, it is often easier to expound the calculations then it is to explain the underlying conceptual structures. But they are there—and now formulated for freshmen in college in this attractive new text, emphasizing the ideas about categories.

Category theory provides a simple and systematic conceptual framework for mathematics including the ideas of sets, functions between sets and of "sets with structure", and the axiomatic formulations of the composition of functions. This text, written by two experts in category theory and tried out carefully in courses at SUNY of Buffalo, provides a simple and effective first course on conceptual mathematics. It offers a careful elementary description of sets and of categories, with many examples—and leading up to many fascinating matters, including the Brouwer fixed point theorem, Cantor's diagonal argument, and the Gödel numbers.

The text starts with the idea of multiplication (for example, Space = Plane $\times$ Line) and shows how this idea leads to Cartesian coordinates and to the description of a product of two objects by diagrams using the projections of the product of two spaces on its factors.

The first example of a category is the category of finite sets and their maps. This is illustrated by both internal diagrams (which element goes where) and by external diagrams (arrows from the domain of a map to the codomain). This leads to the idea of the composition of maps and the associative law for such composition. Here, as throughout the book, the formal definition is followed by many carefully described examples and computations. Also included are typical comments from students and the suggested clarification of their occasional puzzlement. In this way, the book presents an effective dialog with ideas.

From the category of finite sets, the text goes on to the general definition of a category, with its objects, maps between these objects, the composition of these maps, the associative law for composition of maps and the identity maps. Various examples illustrate the importance of the order in which two maps may be composed. In general each section of the book is followed by worked-out exercises, examples of student response, and more exercises.

Special kinds of maps—isomorphisms, the sections and the retracts of a given map, idempotents, injections, etc., are illustrated, defined, and discussed. An informal discussion of continuity indicates that it is plausible that the inclusion map j: $C \to D$ of the boundary circle C of a disc D is not a retract. Using this plausible result, the authors give the standard proof that any continuous map of the disc into itself necessarily has a fixed point (the Brouwer fixed point theorem). Here and elsewhere, the text is careful in the construction and exposition of proofs, so that it provides for the student an effective introduction to the basic idea of "proof".

Part III provides a number of examples of easily accessible categories: the category of sets, the category of sets with an endomorphism (also called the category of automata or of discrete dynamical systems), the category of irreflexive graphs (where each edge is directed from one vertex (dot) to another), the category of reflexive graphs (with an "identity arrow" at each vertex), the category of sets each with an idempotent endomap and the category of arrows of any given category. With these examples, the student is given a view of the idea of a "set with structure" and of structure preserving maps. Again, there are many exercises and comments.

Finite state automata provide a next example of a category. This study of automata leads to examples of cycles of various finite lengths and to the introduction of the natural numbers as the automaton N consisting of the string of natural numbers with the successor map. This example of a "structure" suggests the idea of a map that preserves a given structure. In turn, this leads to the definition of a functor as a map of categories that preserves all the category structure. Then comes the general idea of studying large "objective" categories by functors to them from small "test" categories, with an outlook on the notion of symmetry.

Universals, such as product objects and terminal objects, next appear. A "terminal" object 1 in any category is one such that any object X has a unique or "universal" map $X \to 1$ to this terminal. A product, $X \times Y$ of two objects X and Y with its projections is a diagram

$$X \leftarrow X \times Y \to Y$$

such that any other pair of maps $X \leftarrow A \to Y$ to X and Y can be composed from the two projections and a unique map $A \to X \times Y$. By reversing all the arrows in this definition one has the definition of the "dual" operation of "sum"—or "coproduct". In a category with both products and sums these definitions determine a standard map

$$A \times B + A \times C \to A \times (B + C).$$

The category is said to satisfy the "distributive law" when this standard map is an isomorphism (examples). There is a careful proof of the uniqueness (up to isomorphism) of the product and the sum when they exist.

The fifth chapter introduces the "map objects" Y^T. For example, given sets Y and T, the set Y^T of all the maps f: $T \to Y$ is such a map object. It comes together with the corresponding "evaluation" map

$$Y^T \times T \to Y$$

and the "universal" property of this evaluation. More generally, in any category with products and a terminal object, an object X is said to parametrize all the

maps $T \to Y$ when there is some one map f: $T \times X \to Y$ such that any map $T \to Y$ has the form $f(-, x)$ for some "point" x: $1 \to X$ of X. Then Cantor's diagonal theorem holds in any such category: if an object T has enough points to parametrize all maps $T \to Y$ by means of some map $T \times T \to Y$, then Y has the fixed point property. In particular, the set 2 of two elements does not have the fixed point property. Therefore, for all sets T, one has $T < 2^T$, since 2^T does parametrize all maps $T \to 2$.

After further careful discussion and examples of exponentials, the text turns to the consideration of "parts" (that is, subobjects S) of an object X, and of the characteristic functions h: $X \to 2$ of such subobjects. This characteristic function has the property that the diagram

$$\begin{array}{ccc} S & \longrightarrow & 1 \\ \downarrow & & \downarrow \\ h\colon X & \longrightarrow & 2 = (0,1) \end{array}$$

is a "pullback". But this usual set 2 consisting of the two truth values 0 and 1 must in many categories be replaced by a more sophisticated truth value object. This leads to the definition of a topos: a category with a terminal 1 and initial 0, with products, pullbacks, sums and all exponentials, together with a suitable object t: $1 \to \Omega$ of truth values, such that any part of an object X has a unique characteristic function $X \to \Omega$, much as in the diagram above.

All told, this text on conceptual mathematics thus succeeds in presenting a careful introduction of concepts leading up to non-trivial examples. This reviewer has not yet had occasion to try this text out with students, but he is confident that the care and the wealth of examples will be successful in explaining to students the idea of a proof and the concepts such as "sets with structure" and "structure-preserving" maps. Students are guided to the art of saying things "right". Thus on page 279, Chad is asked

> "To say that T is a terminal object in the category C means what?"
> Chad responds "That there is only one map".
> "One map; from where to where?"
> Chad: "From the other object to T".
> "What other object?"
> Chad: "Any other object".
> "Right. From any other object. So start the sentence with that;
> don't leave it for the end".

In this way Chad and the other students are carefully led to say the things necessary for clarity. The careful use of concepts and their examples through this text leads to precision and thus to understanding.

Department of Mathematics
University of Chicago
5734 University Ave.
Chicago, Illinois 60637-1514
saunders@math.uchicago.edu

Early Astronomy. By Hugh Thurston. Springer-Verlag, 1994; first softcover printing, 1996, x + 268, $29.95.

Reviewed by **Ezra Brown**

Books have always been willing accomplices in my casual but life-long interest in astronomy and outer space. My early interest in stargazing had been aided and abetted by two books. One was *The Stars: A New Way to See Them*, written by H. A. Rey [8] (yes, the author of the "Curious George" books!), which did indeed redraw many of the constellations so as to make them more recognizable to budding young astronomers. The other was Henry M. Neely's fascinating *The Stars by Clock and Fist* [5], which described a method—which I still use and which works beautifully—for locating stars and constellations in which the only thing you needed, besides the book, was an ability to find the North Star. Its author dreamed up the clock-and-fist system while teaching adult education classes in astronomy at the Hayden Planetarium back in the Truman administration.

Since solving mysteries for the uninitiated—in particular, for children—was, in essence, what these books were all about, they had to be well written and clear, or nobody would read them. And they were. Because of them, I could pick up other books on "amateur" astronomy (i.e., any book whose Library of Congress classification number begins with QB63) and make reasonable sense of them by translating right ascensions and declinations, as well as a variety of star charts and maps, into the languages and pictures of Rey and Neely. Furthermore, if a really well-written book on astronomy fell into my hands, it was eagerly read.

And so it was that when Gerald Hawkins' *Stonehenge Decoded* [2] came out many years later, I devoured it. Here was a tantalizing mystery involving astronomy as practiced by people who *really* wanted to understand the sky; so great was their fascination that they created a structure that was part scientific observatory and part sacred space. The mystery was just this: why was Stonehenge built in that manner and in that location, and how was it done? Well, this story is now well known, and a number of Hawkins' deductions about Stonehenge have been subsequently modified. The point is that *Stonehenge Decoded* was, and remains, a well-written book—for adults, even—about a subject of great interest to many lay persons, including myself.

Now, let's jump ahead a decade or so. During the mid-80's, I had acquired a semiprofessional interest in Early Modern Science. After teaching the history of mathematics a few times, I was not happy with the various discussions about the mathematics of the Renaissance. This included, of course, applications of mathematics—in particular, an old friend from the Greek quadrivium, namely astronomy. I began trying to understand how Copernicus might have arrived at the theory that became his *De revolutionibus orbium coelestium*. Subsequent investigations into Copernicus' predecessors led to Ernst Zinner's biography (in German) of that energetic mathematician and astronomer, Johann Müller of Königsberg, called Regiomontanus (1436–1476) [10]. Following a conversation with my friend, historian of science David Lux ("Why not translate it yourself? At a page a day,

you'll be done in a year!"), I did just that. (It was even published—but that's another story.)

This is all by way of explanation that, when Hugh Thurston's *Early Astronomy* arrived in the mail, I could hardly wait to get my hands on it. The first thought that sprang to mind was, "Where was this book when I needed it?" For it would have been—I anticipated—of great help in filling in the multitude of gaps in my knowledge of the details of early astronomy. Such a book, I thought, would describe the instruments used in the pre-telescopic era, explain about oppositions, conjunctions, proper motions, synodic periods, and a host of other astronomical terms, identify the principal problems ancient astronomers were trying to solve, and give some indication of why they were important.

The fairly detailed table of contents looked promising. The book begins with a preliminary chapter of forty-four pages on early stargazers including such standard topics as the rotation of the heavens, the sun, moon, planets and stars, and the tools early astronomers used. The next nine chapters cover the astronomy of several early civilizations, including the megalithic world, Mesopotamia, Egypt, China, Greece, India, the Arabic-speaking Islamic world, the Mayas, and the European Renaissance. About one-fifth of the book is devoted to the achievements of Hipparchus and Ptolemy, and this discussion is supplemented by six appendices. Much would be made clear to me about a subject that my sainted Aunt Mildred would have called a *hegdesh*.

What's a *hegdesh*? Well may you ask. Here's an analogy. You go into an antique shop where you know there are wondrous things to be had, but in order to find them, you have to pick your way through large piles of things stacked here and there. The aisles are narrow, and you know that there's some hint of an orderly plan, but it's hard to figure out just what that plan is. Furthermore, occasional signposts are wrong and others are quite mysterious. Also, in your searching, you have to pay careful attention to details. After a great deal of peeping in here, poking in there, dusting off this, thinking about that, you finally come to some gems. Aunt Mildred would have called this antique shop a *hegdesh*.

Early astronomy is a bit like that. Thurston himself sets the tone of it all on page 4, where he is talking about the constellations: "The shapes of the constellations are fixed. They do not change like the pattern made by a flock of starlings against the sky. But their positions are not fixed: the complete bowl rotates about a fixed point once a day. This point is called the celestial pole." Then he trumps his own ace on page 7, pointing out that "(1) the center of rotation is not fixed; (2) the pattern is not fixed; and (3) one rotation does not take exactly one day."

In a nutshell, this was the essential problem facing early astronomers. Celestial objects appear to move in nice, simple, orderly patterns. But they do not. The total motion of a celestial object is mainly regular, but with small variations, not obvious to the casual eye, that become more discernible with long-term observations: general regularities tempered with small irregularities. Knowing this prompts many questions. What tools did the early astronomers use to measure time, directions, and motion? At what point did they begin to use mathematics to make some attempt at predicting future positions of celestial objects? In short, How Did They Figure It All Out?

Alas, this book is not an easy read—in some places, it is very slow going—and so it became clear that this would not help make early astronomy less of a *hegdesh*. One difficulty is that Thurston's treatment of his subject matter is very understated. Another is that his writing style is more descriptive than compelling. Curious about the author's motivation and purpose for writing the book, I turned

to the introduction, which states that "This book covers astronomy from the beginning up to the time of Kepler. Most of this was developed at a time when other sciences, notably physics and chemistry, had scarcely started. The scientific edifice built up over this period is one of the triumphs of human intellect." There is no statement of purpose or "why this book came to be," other than to *cover* the material. And nothing turns students off more than professors who just "cover the material."

Take the topic of eclipses, for example. There is all of the information necessary to understand the motions of the sun and the moon and to understand why eclipses seem to occur at approximately regular intervals, but it is not written in a way designed to excite or entice the reader. Another example is the treatment of Stonehenge. Thurston ends his discussion on Stonehenge by pointing out that the alignments there were set up using only very painstaking and reasonably accurate observations—as opposed to advanced mathematical or astronomical theories—and states that the marvel of the place is not in its astronomy but in its construction. But Hawkins' book on the same subject conveys a great deal of wonder both at the construction *and* at the observations. The wonder also lay in figuring out that yes, the structure's alignments *are* astronomical in character, and this comes through in Hawkins' writing very clearly. For Thurston, the wonder apparently disappears once you see "how the trick is done."

In addition, there are some occurrences in the book that are somewhat distracting. One that turns out to be amusing is in a passage on the orientation of the pyramids: "To find north or south accurately we must use the stars; the sun is too big and too bright to yield an accurate result. I describe one way in which north can be found on page 26." But this passage *is* on page 26! (After a number of false starts and failed searches, I found the reference close to the *end* of page 26.) Not so amusing is a passage on Hipparchus and the length of the year. In one sentence, it is written that "the time interval between the summer solstices of 1990 and 1991 was 365.2403 days, whereas between the summer solstices of 1991 and 1992 it was 365.2465 days." A couple of sentences later, we read that "the year A.D. 1 was 365.242187 days long and the year decreases by 0.000006 days per century." These two sentences seem to contradict one another. A third example is this: although there is a great deal of discussion about large-scale chronology—the different kinds of years and months—there are almost no statements about how these early astronomers measured the passage of time during a day. It is mentioned that the Babylonians timed the culmination (highest point) of one star by recording which other star is rising, but that is about it for "telling time." Finally, I was disappointed that Regiomontanus does not appear in the book (I'll get over it).

But let us return to the writing style. Since there is a great deal of terminology to absorb, it might be the case that something intrinsic in the subject of ancient astronomy lends itself to tediosity (if there is no such word, there should be). So, I looked through several other books on the subject.

I began with Neugebauer's *The Exact Sciences in Antiquity* **[6]** and van der Waerden's *Science Awakening II: The Birth of Astronomy* **[9]**, just to make sure. They include as much detail as does Thurston (maybe more!), but they have been written in such a way as to make the subject matter compelling. They set the hook and lure the reader on, giving manageable bites of technical information on which to chew.

I also read through parts of Neugebauer's monumental three-volume treatise, *A History of Ancient Mathematical Astronomy* **[7]**, which is both unapologetically

mathematical—not for the faint of heart or the "trigonometrically challenged," and fairly slow going even for an interested mathematician—and an absorbing read. This treatment includes lots of illustrative examples worked out in detail, so that the exposition is straightforward to follow.

Yet another path to early astronomy is the field of archaeoastronomy—the interdisciplinary study of ancient, prehistoric, and traditional astronomy and its cultural context. If you are interested in this point of view, then E. C. Krupp's *Echoes of the Ancient Skies* [4] is a treat. This book is chock-full of information about ancient and prehistoric observatories, myths, calendars, ceremonies and cosmologies from around the world, but it also contains enough of the essentials on how the heavens move to interest any reader who is a hard-core astronomy buff. Krupp is clearly caught up in the wonder and the beauty of it all, and his writing shows it. Other books along the same line are two collections, *Astronomy of the Ancients*, edited by K. Brecher and M. Feirtag [1], and *In Search of Ancient Astronomies* [3], edited by Krupp.

What it looks like is that Thurston's book is "neither fish nor fowl nor good red herring": it's not strictly (1) mathematical, (2) historical, or (3) mystical/religious/interdisciplinary. Also, its writing style is not very compelling. But it contains the information—it covers the subject. Also, the gems are there for the polishing, but you might have to work at it. (One of the gems is almost at the end of the book: Thurston's description of Kepler's efforts at finding the orbit of Mars. Would that the entire book were that intriguing!)

Oh, yes. You may have been wondering, "Just what is Neely's clock-and-fist method?" Easily explained. Instead of using compass directions to tell you which way to face in order to see a particular constellation, you use the clock directions with 12 o'clock being north (toward the North Star). And instead of using a sextant to measure angles up from the horizon, you use your fist to get a line-of-sight to the horizon, and then sight upward however many fists you need to get to the particular constellation. Here, the convention is that from horizon to zenith equals nine fists. And now, go and explore the night sky yourself. Have fun... there are great wonders out there to behold!

REFERENCES

1. Kenneth Brecher and Michael Feirtag, eds., *Astronomy of the Ancients*. MIT Press, Cambridge, Massachusetts, 1979.
2. Gerald Hawkins, *Stonehenge Decoded*. Doubleday, New York, 1965.
3. E. C. Krupp, ed., *In Search of Ancient Astronomies*. McGraw-Hill, New York, 1979.
4. E. C. Krupp, *Echoes of the Ancient Skies*. Harper & Row, New York, 1983; Oxford University Press, New York, 1994.
5. Henry M. Neely, *The Stars by Clock and Fist*. The Viking Press, New York, 1956.
6. Otto Neugebauer, *The Exact Sciences in Antiquity*. Brown University Press, Providence, 1957; Dover Publications, New York, 1969.
7. Otto Neugebauer, *A History of Ancient Mathematical Astronomy*, 3 vols. Springer-Verlag, New York, 1975.
8. H. A. Rey, *The Stars: A New Way to See Them*. Houghton Mifflin, Boston, 1980.
9. B. van der Waerden, *Science Awakening II: The Birth of Astronomy*. Oxford University Press, New York, 1974.
10. Ernst Zinner, *Regiomontanus: His Life and Work*, translated by Ezra Brown. North-Holland, Amsterdam, 1990.

Department of Mathematics
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061
brown@math.vt.edu

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1–4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, T*(14: 1), S*, L*. *Laboratories in Mathematical Experimentation: A Bridge to Higher Mathematics.* Mount Holyoke College. Springer-Verlag, 1997, xix + 278 pp, $34.95 (P). [ISBN 0-387-94922-4] Sixteen, basically independent, labs that help students understand the experimental nature of mathematics before they delve into theoretical courses. Access to computers is required; knowledge of calculus is not. An excellent resource. CEC

General, P, L*.** *Indiscrete Thoughts.* Gian-Carlo Rota. Ed: Fabrizio Palombi. Birkhäuser Boston, 1997, xxii + 280 pp, $36.50. [ISBN 0-8176-3866-0] In wide-ranging essays, memoirs, book reviews, and other assorted genres, Rota surveys nearly 50 years of life in academic mathematics and philosophy. The menu includes social commentary, mathematical gossip, academic philosophy, advice to the young ("Ten Lessons I Wish I had Learned"), and much more. Learned, thought-provoking, politically incorrect, delighting in paradox, and likely to offend—but everywhere readable and entertaining. PZ

General, P. *The Legacy of Norbert Wiener: A Centennial Symposium.* Eds: David Jerison, I.M. Singer, Daniel W. Stroock. Proc. of Symp. in Pure Math., V. 60. AMS, 1997, xix + 405 pp, $80. [ISBN 0-8218-0415-4] Proceedings of a 1994 event at M.I.T.

Reference, P, L. *Handbook of Mathematics.* I.N. Bronshtein, K.A. Semendyayev. Transl: K.A. Hirsch. Springer-Verlag, 1997, xv + 973 pp, $57. [ISBN 3-540-62130-X] Reprint of the revised third edition (TR, October 1986).

Recreational Mathematics, P, L. *The Green Book of Mathematical Problems.* Kenneth Hardy, Kenneth S. Williams. Dover, 1997, ix + 173 pp, $6.95 (P). [ISBN 0-486-69573-5] Republication, with corrections, of *The Green Book: 100 Practice Problems for Undergraduate Mathematics Competitions* published by Integer Press in 1985 (TR, March 1986).

Recreational Mathematics, P, L*. *Penrose Tiles to Trapdoor Ciphers . . . and the Return of Dr. Matrix.* Martin Gardner. MAA, 1997, ix + 319 pp, $27.95 (P). [ISBN 0-88385-521-6] This revised edition features a new bibliography, corrections to the text, and a postscript written by the author. (W.H. Freeman edition, TR, March 1989.)

Education, P, L. *Bold Ventures, Volume 1.* Eds: Senta A. Raizen, Edward D. Britton. Kluwer Academic, 1997, xviii + 249 pp, $125. [ISBN 0-7923-4231-3] Overview and synthesis of innovations in U.S. mathematics and science education based on eight case studies (five in science, three in mathematics) carried out over the past five years. Covers the context and motivation of reform; changing conceptions of science, mathematics, and instruction; changing roles of teachers; and "underplayed" issues (assessment, evaluation, equity, and diversity). LAS

History, P. *Helmut Wielandt: Mathematische Werke, Mathematical Works. Volume 2: Linear Algebra and Analysis.* Eds: Bertram Huppert, Hans Schneider. Walter de Gruyter, 1996, xx + 802 pp, DM 348. [ISBN 3-11-012453-X]

Logic, P. *Introduction to Mathematical Logic.* Alonzo Church. Landmarks in Math. &

Physics. Princeton Univ Pr, 1996, ix + 378 pp, \$19.95 (P). [ISBN 0-691-02906-7] Republication of the 1958 corrected second printing.

Combinatorics, P. *Probabilistic and Analytical Aspects of the Umbral Calculus.* A. Di Bucchianico. CWI Tract, V. 119. Centrum voor Wiskunde en Informatica, 1997, 148 pp, Dfl. 35 (P). [ISBN 90-6196-471-7]

Discrete Mathematics, P. *Operations Research and Discrete Analysis.* Alekseĭ D. Korshunov. Math. & Its Applic., V. 391. Kluwer Academic, 1997, vii + 331 pp, \$175. [ISBN 0-7923-4334-4] Translations of papers from the second volume of the Russian-language journal *Diskretnyĭ i Analiz i Issledovanie Operaatsiĭ.*

Number Theory, P. *Primes of the Form $x^2 + ny^2$: Fermat, Class Field Theory, and Complex Multiplication.* David A. Cox. Pure & Appl. Math. Wiley, 1989, xi + 351 pp, \$49.95 (P). [ISBN 0-471-19079-9] Paperback republication (TR, December 1990).

Number Theory, P. *Continued Fractions.* A. Ya. Khinchin. Dover, 1997, xi + 95 pp, \$6.95 (P). [ISBN 0-486-69630-8] Republication of the English translation of the third Russian edition (1961) published by the U. of Chicago in 1964.

Number Theory, S(15–16), P, L*. *The Book of Numbers.* John H. Conway, Richard K. Guy. Springer-Verlag, 1996, ix + 310 pp, \$29. [ISBN 0-387-97993-X] A marvelous compendium of fascinating facts pertaining to numbers, from the natural to the surreal, by two masters of the field. BC

Number Theory, T(16–17: 2).** *Elliptic Functions: A Constructive Approach.* Peter L. Walker. Wiley, 1996, xv + 214 pp, \$62.95. [ISBN 0-471-96531-6] A thoroughly modern introduction to the theory of elliptic functions suitable for use with (strong) undergraduates. Assumes only basic topology and a bit of complex analysis. Beginning with Eisenstein series (where else?), the author carefully constructs basic elliptic functions, theta functions, Jacobian functions, elliptic integrals, and modular functions. A perfect starter book for anyone wanting to work through Wiles' proof of Fermat's Last Theorem. MPR

Linear Algebra, T*(14: 1), C. *Interactive Linear Algebra: A Laboratory Course Using Mathcad.* Gerald J. Porter, David R. Hill. Springer-Verlag, 1996, \$42.95 (P), with disks. [ISBN 0-387-94608-X] Uses Mathcad (5.0 or 6.0) to create a discovery-based laboratory learning environment. Covers the usual topics (eigenvalues, linear transformations, and vector spaces, etc.). Exercises ask for written summaries of the mathematics. The text is a printout of the electronic book. JNC

Group Theory, P. *Group Theory in China.* Eds: Zhe-xian Wan, Sheng-ming Shi. Math. & Its Applic. Kluwer Academic, 1996, 261 pp, \$120. [ISBN 0-7923-3989-4] 15 essays by former students and colleagues of Hsio-Fu Tuan.

Algebra, S(15), L. *Applied Abstract Algebra.* Ed: S.K. Jain. Centre for Professional Development in Higher Education (Univ. of Delhi, Delhi–110 007, India), 1996, 105 pp, (P). Notes from a workshop at the University of Delhi. Deals with standard applications such as block designs, Burnside's and Pólya's theorems, switching circuits, matroids, and cryptography. Includes references and occasional exercises. CEC

Algebra, P. *Ordered Algebraic Structures.* Eds: W. Charles Holland, Jorge Martinez. Kluwer Academic, 1997, ix + 332 pp, \$162. [ISBN 0-7923-4377-8] Proceedings of the June 1995 Curaçao conference.

Algebra, P. *Foundations of Lie Theory and Lie Transformation Groups.* V.V. Gorbatsevich, A.L. Onishchik, E.B. Vinberg. Springer-Verlag, 1997, 235 pp, \$54.50 (P). [ISBN 3-540-61222-X] Republication of *Lie Groups and Lie Algebras I* (V. 20 of the Encyclopeaedia of Mathematical Sciences).

Algebra, P. *Basic Notions of Algebra.* I.R. Shafarevich. Springer-Verlag, 1997, 258 pp, \$54.50 (P). [ISBN 3-540-61221-1] Republication of *Algebra I* (V. 11 of the Encyclopeaedia of Mathematical Sciences).

Algebra, P. *Finite Fields and Applications.* Eds: S. Cohen, H. Niederreiter. London Math. Soc. Lect. Note Ser., V. 233. Cambridge Univ Pr, 1996, xx + 401 pp, \$42.95 (P). [ISBN 0-521-56736-X] Proceedings of a 1995 conference in Glasgow. 27 survey and research papers on theory and applications.

Complex Analysis, T(17: 2). *Complex Variables: Introduction and Applications.* Mark J. Ablowitz, Athanassios S. Fokas. Texts in Appl. Math. Cambridge Univ Pr, 1997, xii + 647 pp, \$34.95 (P); \$69.95. [ISBN 0-521-48523-1; 0-521-48058-2] An encyclopedic treatment. In two parts: an introduction with the elements typically found in a first course through residue theory and conformal maps; second part covers advanced topics including asymptotic evaluation of integrals and Riemann–Hilbert problems. Plenty of well-chosen exercises. TAV

Complex Analysis, T(17: 2). *Function Theory*

of One Complex Variable. Robert E. Greene, Steven G. Krantz. Pure & Appl. Math. Wiley, 1997, xiii + 496 pp, $69.95. [ISBN 0-471-80468-1] Emphasizes connections with and differences from the results of multivariable calculus. Very readable. Extensive problem sets. TAV

Complex Analysis, T(18: 2), S, P. *Compact Riemann Surfaces: An Introduction to Contemporary Mathematics.* Jürgen Jost. Transl: R.R. Simha. Universitext. Springer-Verlag, 1997, xiv + 291 pp, $49.95 (P). [ISBN 3-540-53334-6] An introduction, but unusual in its breadth—the author aims to exhibit connections between Riemann surfaces and many other areas, including differential geometry, algebraic topology, algebraic geometry, calculus of variations, and elliptic differential equations. Also introduces Teichmüller theory. Treatment of Riemann surfaces focuses on three fundamental theorems: the uniformization theorem, Teichmüller's theorem, and the Riemann–Roch theorem. Contains occasional helpful figures and a generous collection of exercises. PZ

Differential Equations, S, C, P.** *VisualDSolve: Visualizing Differential Equations with Mathematica.* Dan Schwalbe, Stan Wagon. Springer-Verlag, 1997, xiv + 271 pp, $34.95 (P), with disk. [ISBN 0-387-94721-3] Mathematica tools for visualizing solutions of differential equations. First part (5 chapters) is a user's manual. Second part (12 chapters) illustrates use of these tools in various modeling contexts.

Differential Equations, T(17:1), P. *Ordinary Differential Equations in the Complex Domain.* Einar Hille. Dover, 1997, xi + 484 pp, $14.95 (P). [ISBN 0-486-69620-0] Republication of the 1976 Wiley edition (TR, November 1976).

Partial Differential Equations, P. *Solitons, Geometry, and Topology: On the Crossroad.* Eds: V.M. Buchstaber, S.P. Novikov. AMS Transl., Ser. 2, V. 179. AMS, 1997, x + 189 pp, $89. [ISBN 0-8218-0666-1] 9 papers written by participants in the S.P. Novikov seminar on topology and mathematical physics in Moscow.

Partial Differential Equations, T(15: 2). *Partial Differential Equations and Mathematica.* Prem K. Kythe, Pratap Puri, Michael R. Schäferkotter. CRC Pr, 1997, xx + 378 pp, $59.95. [ISBN 0-8493-7853-2] For a first course in PDEs. Covers basic concepts and methods. Uses *Mathematica* throughout to illustrate ideas and techniques. AO

Dynamical Systems, P. *Dynamical Systems in Cosmology.* Eds: J. Wainwright, G.F.R. Ellis. Cambridge Univ Pr, 1997, xiv + 343 pp, $69.95. [ISBN 0-521-55457-8] 15 articles outline recent progress in analyzing the evolution and structure of cosmological models from a dynamical systems viewpoint and relate this approach to others.

Dynamical Systems, P. *Ordinary Differential Equations and Smooth Dynamical Systems.* D.V. Anosov, *et al.* Springer-Verlag, 1997, 233 pp, $54.50 (P). [ISBN 3-540-61220-3] Republication of *Dynamical Systems I* (V. 1 of the Encyclopeadedia of Mathematical Sciences). (TR, April 1989.)

Dynamical Systems, P. *Chaotic Dynamics of Nonlinear Systems.* S. Neil Rasband. Wiley, 1997, x + 230 pp, $42.95 (P). [ISBN 0-471-18434-9] Paperback edition of 1990 text (TR, December 1990).

Dynamical Systems, P. *Integrable Systems and Foliations: Feuilletages et Systèmes Intégrables.* Eds: Claude Albert, Robert Brouzet, Jean Paul Dufour. Progress in Math., V. 145. Birkhäuser Boston, 1997, x + 212 pp, $69.50. [ISBN 0-8176-3894-6] 11 papers from a 1995 colloquium held in Montpellier, France to honor Pierre Molino.

Numerical Analysis, P, L. *Parallel Numerical Algorithms.* Eds: David E. Keyes, Ahmed Sameh, V. Venkatakrishnan. ICASE/LaRC Interdisc. Ser. in Sci. & Eng., V. 4. Kluwer Academic, 1997, xi + 395 pp, $195. [ISBN 0-7923-4282-8] 14 survey papers on large-scale scientific computing from a 1994 ICASE/NASA workshop. Sections: Linear Systems; Preconditioning; "Fast" Application of Operators; Parallel Tools, Environments, and Benchmarking.

Numerical Analysis, T(17: 2), P, L. *Matrix Computations, Third Edition.* Gene H. Golub, Charles F. Van Loan. Johns Hopkins Univ Pr, 1996, xxvii + 694 pp, $29.95 (P); $65. [ISBN 0-8018-5414-8; 0-8018-5413-X] Revised and updated edition (*Second Edition*, TR, March 1990). Includes new material on methods for sparse unsymmetric linear systems. AO

Numerical Analysis, P. *Polynomial Based Iteration Methods for Symmetric Linear Systems.* Bernd Fischer. Wiley, 1996, 283 pp, $49.95. [ISBN 0-471-96796-3] An overview of the state-of-the-art. Emphasizes properties of the underlying polynomials rather than matrix manipulations. Includes MATLAB code for the various algorithms discussed. AO

Operator Theory, P. *Hardy Classes and Operator Theory.* Marvin Rosenblum, James Rovnyak. Dover, 1997, xii + 161 pp, $9.95 (P).

[ISBN 0-486-69536-0] Republication, with corrections, of the Oxford University Press edition (TR, November 1986).

Operator Theory, P. *Linear and Nonlinear Perturbations of the Operator* div. V.G. Osmolovskiĭ. Transl. of Math. Mono., V. 160. AMS, 1997, xiii + 104 pp, $59. [ISBN 0-8218-0586-X]

Functional Analysis, P. *Multidimensional Complex Analysis and Partial Differential Equations.* Eds: Paulo D. Cordaro, Howard Jacobowitz. Contemp. Math., V. 205. AMS, 1997, ix + 276 pp, $55 (P). [ISBN 0-8218-0509-6] Proceedings of a 1995 conference held in São Carlos, Brazil, to honor François Treves.

Analysis, T(17: 1), P. *Symbolic Integration I: Transcendental Functions.* Manuel Bronstein. Algorithms & Comput. in Math., V. 1. Springer-Verlag, 1997, xiii + 299 pp, $49. [ISBN 3-540-60521-5] Detailed presentation of a version of the Risch algorithm for finding symbolic and antiderivatives. Because the book is restricted to integration of transcendental functions, the basic mathematical prerequisite is a first course in abstract algebra (i.e., rings, fields, and polynomials). AO

Analysis, T(15–16: 3), L. *Undergraduate Analysis, Second Edition.* Serge Lang. Undergrad. Texts in Math. Springer-Verlag, 1997, xv + 642 pp, $54.95. [ISBN 0-387-94841-4] A broad and deep introduction to many areas in advanced undergraduate analysis, ranging as far as rudiments of algebraic topology, functional analysis, and differential equations. Assumes thorough grounding in analysis and linear algebra, and considerable mathematical maturity. Clear, brisk, no-nonsense expository style; many exercises, relatively few figures. Main content change from previous edition (TR, November 1985) is a new chapter on locally integrable vector fields. PZ

Analysis, P. *Asymptotics and Special Functions.* Frank W.J. Olver. AK Peters, 1997, xviii + 572 pp, $69. [ISBN 1-56881-069-5] Republication, with corrections, of the Academic Press edition (TR, November 1974).

Analysis, P. *Harmonic Functions on Trees and Buildings.* Ed: Adam Korányi. Contemp. Math., V. 206. AMS, 1997, x + 181 pp, $35 (P). [ISBN 0-8218-0605-X] Proceedings of a 1995 workshop held in New York City. Part I contains 4 expository papers. Part II contains extended abstracts of 19 papers given at the workshop.

Geometry, P. *Lectures in Real Geometry.* Ed: Fabrizio Broglia. Expos. in Math., V. 23. Walter de Gruyter, 1996, xiv + 268 pp, DM 198. [ISBN 3-11-015095-6] Expanded versions of lectures given at the 1994 Winter School in Real Geometry at the Universidad Complutense de Madrid.

Geometry, P. *The Arnold–Gelfand Mathematical Seminars: Geometry and Singularity Theory.* Eds: V.I. Arnold, *et al.* Birkhäuser Boston, 1997, vii + 437 pp, $89.50. [ISBN 0-8176-3883-0]

Geometry, P. *Finite Geometries.* Peter Dembowski. Classics in Math. Springer-Verlag, 1997, xi + 375 pp, $35 (P). [ISBN 3-540-61786-8] Paperback edition of the 1968 text (TR, January 1969).

Topology, T(17), L. *Three-Dimensional Geometry and Topology, Volume 1.* William P. Thurston. Math. Ser., V. 35. Princeton Univ Pr, 1997, x + 311 pp, $39.50. [ISBN 0-691-08304-5] Extremely well-written, highly intuitive treatment of hyperbolic geometry of 3-manifolds. Thurston succeeds in motivating his subject while at the same time providing rigor. JD

Operations Research, T(16: 2), C, L. *Operations Research: An Introduction, Sixth Edition.* Hamdy A. Taha. Prentice Hall, 1997, xx + 916 pp, with disk. [ISBN 0-13-272915-6] Substantially revised and updated. Balances model formulation, computation, and theory. Uses numerical examples to introduce ideas. Accompanying diskette includes IBM PC software. (*Second Edition*, Extended Review, March 1980.) AO

Operations Research, P. *Frontiers in Queueing: Models and Applications in Science and Engineering.* Ed: Jewgeni H. Dshalalow. Prob. & Stoch. Ser. CRC Pr, 1997, 468 pp, $89.95. [ISBN 0-8493-8076-6] Engineering-oriented surveys with discussions of open problems and future research directions. 16 chapters in 4 sections: progress of classical queueing methods; telecommunications and computer networks; traffic processes; applied techniques and statistical inference in queueing models.

Optimization, P. *Convex Analysis.* R. Tyrrell Rockafellar. Landmarks in Math. & Physics. Princeton Univ Pr, 1997, xviii + 451 pp, $22.95 (P). [ISBN 0-691-01586-4] Paperpack edition of this classic (TR, April 1970).

Optimization, T(18: 1), P, L. *Primal-Dual Interior-Point Methods.* Stephen J. Wright. SIAM, 1997, xx + 289 pp, $37 (P). [ISBN 0-89871-382-X] Primal-dual methods are the most important and useful interior point algo-

rithms for solving linear programming problems. Text presents the theoretical properties of these algorithms, discusses their practical implementation, and provides information about current software. AO

Probability, T?(16–17: 2), P*, L*. *Probability Theory: Collection of Problems.* A. Ya. Dorogovtsev, *et al.* Transl. of Math. Mono., V. 163. AMS, 1997, xi + 347 pp, $119. [ISBN 0-8218-0372-7] An immense collection of probability problems ranging from the simplest combinatorial ones to limit theorems for Wiener processes. An invaluable tool for the teacher; an intriguing choice for a problems-based course. Provides answers or hints for most problems. TAV

Stochastic Processes, T(18: 1), P. *Discrete Gambling and Stochastic Games.* Ashok P. Maitra, William D. Sudderth. Applic. of Math., V. 32. Springer-Verlag, 1996, xi + 244 pp, $49.95. [ISBN 0-387-94628-4] The authors write: "These methods [maximizing chances of beating a casino or winning against an intelligent opponent] are at the heart of modern theory of stochastic control and stochastic games." Taking up from Dubins and Savages' classic, but limiting their attention to countably additive spaces, the authors summarize and extend the theory of optimization in a gambling environment. Pays attention to the differences between leavable and nonleavable games. Presumes a course in measure theoretic probability. TAV

Stochastic Processes, P. *Elements of the Theory of Markov Processes and Their Applications.* A.T. Bharucha-Reid. Dover, 1997, xi + 468 pp, $14.95 (P). [ISBN 0-486-69539-5] Republication of 1960 McGraw-Hill edition.

Stochastic Processes, P. *Introduction to the Theory of Random Processes.* I.I. Gikhman, A.V. Skorokhod. Transl: Scripta Technica, Inc. Dover, 1996, xiii + 516 pp, $14.95 (P). [ISBN 0-486-69387-2] Republication of the 1969 W.B. Saunders edition (TR, January 1970).

Stochastic Processes, P. *Proceedings of the Norbert Wiener Centenary Congress, 1994.* Eds: V. Mandrekar, P.R. Masani. Proc. of Symp. in Appl. Math., V. 52. AMS, 1997, xlviii + 566 pp, $99. [ISBN 0-8218-0452-9]

Elementary Statistics, T(13-14: 2). *Statistics and Probability in Modern Life, Sixth Edition.* Joseph Newmark. Saunders College, 1997, xix + 827 pp, $43; $33.75 (P). [ISBN 0-03-006393-0] Updated examples and exercises with sources clearly indicated (*Fifth Edition*, TR, January 1993). Chapters begin with newspaper clippings that set the stage and historical context. MINITAB commands, methodology, and applications to various disciplines integrated throughout. RS

Statistical Methods, P. *MSI–2000: Multivariate Statistical Analysis in Honor of Professor Minoru Siotani on his 70th Birthday.* Eds: Takesi Hayakawa, Makoto Aoshimna, Kunio Shimizu. Amer. Journ. of Math. & Management Sci., V. 16, Nos. 1 & 2. 1996, 266 pp, $125 (P). [ISBN 0-935950-39-7] Proceedings of a 1995 conference at the University of Hawaii.

Statistical Methods, P. *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling.* Ed: Sabine Van Huffel. SIAM, 1997, xxii + 377 pp, $56 (P). [ISBN 0-89871-393-5] Proceedings of the Second International Workshop on Total Least Squares and Errors-in-Variables Modeling held in Leuven, Belgium in August 1996.

Statistical Methods, T(18), P. *Methods for Statistical Data Analysis of Multivariate Observations, Second Edition.* R. Gnanadesikan. Ser. in Prob. & Stat. Wiley, 1997, xvi + 353 pp, $69.95. [ISBN 0-471-16119-5] Expanded coverage and new methods (*First Edition*, TR, November 1977). No exercises. Appendix provides advice on software and lists functions available in S-Plus and SAS. RS

Statistical Methods, T(15–18), P. *Handbook of Parametric and Nonparametric Statistical Procedures.* David J. Sheskin. CRC Pr, 1997, 719 pp, $69.95. [ISBN 0-8493-3119-6] Assumes basic familiarity with descriptive statistics and experimental design (reviewed in an introductory section). Emphasizes applications over theory. Provides practical guidelines for selecting appropriate tests for evaluating a design. Well-organized and thorough. Discusses related tests for each procedure. RS

Statistical Methods, T(18), P. *Survey Measurement and Process Quality.* Eds: Lars Lyberg, *et al.* Ser. in Prob. & Stat. Wiley, 1997, xvii + 777 pp, $79.95. [ISBN 0-471-16559-X] Contributions from various authors are organized into the following sections: Questionnaire Design; Data Collection; Post Survey Processing and Operations; Quality Assessment and Control; Error Effects on Estimation, Analyses, and Interpretation. RS

Statistical Methods, T(14–16). *Survival Analysis: A Self-Learning Text.* Stat. in Health Sci. Springer-Verlag, 1996, xii + 324 pp, $44.95. [ISBN 0-387-94543-1] Gentle, but powerful, introduction to techniques for analyzing censored data. Each chapter begins with an abbreviated outline and a list of objectives, and

concludes with a "test" over the material. Exposition is clear and graphics are insightful. Numerous exercises and data sets. MK

Statistics, P*. *Discrete Multivariate Distributions.* Norman L. Johnson, Samuel Kotz, N. Balakrishnan. Ser. in Prob. & Stat. Wiley, 1997, xxii + 299 pp, $79.95. [ISBN 0-471-12844-9] In 1969 Johnson and Katz wrote *Distributions in Statistics*, the definitive work in the area at the time. This is the fourth volume of their reworking of that book. Originally only one chapter, this book reflects the immense amount of development in the area. After three chapters of introduction to the general topic, separate chapters fully develop the properties of the most important discrete multivariate distributions. TAV

Mathematical Computing, C. *Programming in Mathematica, Third Edition.* Roman E. Maeder. Addison-Wesley, 1997, xvi + 366 pp, $34.38 (P). [ISBN 0-201-85449-X] Updated and expanded to suit Version 3. A thorough and systematic introduction, aided by countless examples, large and small. Large-scale packages—with their attendant software engineering issues—are a special emphasis. PZ

Mathematical Computing, P. *Numerical Methods and Software Tools in Industrial Mathematics.* Eds: Morten Dæhlen, Aslak Tveito. Birkhäuser Boston, 1997, 400 pp, $69.95. [ISBN 0-8176-3973-X] 19 articles present research results from projects funded by the Research Council of Norway between 1992 and 1996. 3 parts: numerical software tools; partial differential equations; geometric modeling.

Applications (Biological Science), T(15-17: 1), L*. *Population Biology: Concepts and Models.* Alan Hastings. Springer-Verlag, 1997, xvi + 220 pp, $49.95. [ISBN 0-387-94862-7] Mathematical models for single and interacting species. Integrates biology and mathematics of ecology. Only mathematical prerequisite is one year of calculus.

Applications (Fluid Mechanics), P. *Annual Review of Fluid Mechanics, Volume 29, 1997.* Eds: John L. Lumley, Milton van Dyke, Helen L. Reed. Annual Reviews, 1997, 630 pp, $60. [ISBN 0-8243-0729-1]

Applications (Physical Science), P. *Seismic Waves in Laterally Inhomogeneous Media, Part II.* Eds: Ivan Pšenčík, Vlastislav Červený, Luděk Klimeš. Birkhäuser Boston, 1996, 369 pp, $34.50 (P). [ISBN 0-8176-5651-0] 13 papers from a 1995 workshop in the Czech Republic. (Also published as Nos. 3 & 4 of V. 148 of *Pure and Applied Geophysics.*)

Applications (Physical Science), P. *Mathematical Aspects of Classical and Celestial Mechanics, Second Edition.* V.I. Arnold, V.V. Kozlov, A.I. Neishtadt. Springer-Verlag, 1997, xiv + 291 pp, $54.50 (P). [ISBN 3-540-61224-6] Republication of the 1993 second edition.

Applications (Physics), S(18), L. *Constitutions of Matter: Mathematical Modeling the Most Everyday of Physical Phenomena.* Martin H. Krieger. Univ of Chicago Pr, 1996, xxii + 343 pp, $65. [ISBN 0-226-45304-9] An intriguing account of one of the central problems of modern physics, namely the mathematical description of matter. Using thermodynamics and statistical mechanics (specifically the Ising Model), the author presents the argument that the mathematics is not only a useful tool, but has true physical meaning. Far from another "physics for the masses" book. This is a solid, scholarly (and lively) attempt to fuse mathematics, physics, and philosophy. MPR

Applications (Systems Theory), P. *Mathematics of Stochastic Manufacturing Systems.* Eds: G. George Yin, Qing Zhang. Lect. in Appl. Math., V. 33. AMS, 1997, xii + 399 pp, $69 (P). [ISBN 0-8218-0755-2] Proceedings of a 1996 AMS-SIAM Summer Seminar held at the College of William and Mary.

Applications, T(15–16), S. *Compliance Quantified: An Introduction to Data Verification.* Rudolf Avenhaus, Morton John Canty. Cambridge Univ Pr, 1996, xiii + 256 pp, $59.95. [ISBN 0-521-55366-0] Covers theory of verification and strategies for implementation. Assumes calculus and basic probability/statistics background. Theory discussed using real-world examples/data from treaty verification and environmental monitoring. No exercises, but a good classroom text nevertheless. MK

Applications, P. *Numerical Methods in Smog Prediction.* M. van Loon. CWI Tract, V. 120. Stichting Mathematisch Centrum, 1997, iv + 149 pp, Dfl. 35 (P). [ISBN 90-6196-473-3]

Applications, P. *Multidisciplinary Design Optimization: State of the Art.* Eds: Natalia M. Alexandrov, M.Y. Hussaini. SIAM, 1997, xvi + 455 pp, $61 (P). [ISBN 0-89871-359-5] Proceedings of a 1995 ICASE/NASA Langley workshop.

Reviewers

JNC: Judith N. Cederberg, St. Olaf; BC: Barry Cipra, St. Olaf; CEC: Clifton E. Corzatt, St. Olaf; JD: Jill Dietz, St. Olaf; MK: Michael Kahn, St. Olaf; AO: Arnold Ostebee, St. Olaf; MPR: Matthew P. Richey, St. Olaf; RS: Richard Single, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf; TAV: Theodore A. Vessey, St. Olaf; PZ: Paul Zorn, St. Olaf.

THE AUTHORS

BRUCE BERNDT'S destiny can be traced back to December 22, 1858, the birth date of his great-grandfather Friedrich Griegoleit in Prussia. He then chose to be born on a prime numbered day in a prime numbered month in a year that (mod 100) is the product of these same two primes. He learned numbers at the age of 1 or 2 by watching his father play pinochle and poker each Sunday evening at his grandmother's home. In the past 30 years he has been nurtured at the University of Illinois by a wealth of outstanding colleagues and graduate students in number theory, two of whom are coauthors of this paper. His fifth and final volume on Ramanujan's notebooks was published by Springer-Verlag near the end of 1997.

HENG HUAT CHAN received his B.S. degree from the National University of Singapore in 1991 and completed his Ph.D. degree under the supervision of Bruce C. Berndt at the University of Illinois at Urbana-Champaign. From September 1995 to June 1996, he was a research scholar at the Institute for Advanced Study (Princeton). During the 1996–97 academic year, he was a visiting professor at the National Chung Cheng University in Taiwan. In August he returned to the National University of Singapore to accept a tenure-track position.

LIANG-CHENG ZHANG is an associate professor of mathematics at Southwest Missouri State University. He received B.S. and M.S. degrees in mathematics from Peking (Beijing) University and a Ph.D. from the University of Illinois. His main research interests are in number theory. He likes to play tennis, ping-pong, and go.

KEN ONO is an assistant professor in the Department of Mathematics at the Penn State University. He received his Ph.D. under the direction of Professor Basil Gordon at UCLA in 1993. Since then he has held post-doctoral positions at the University of Georgia (1993–94), University of Illinois (1994–95), and the Institute for Advanced Study (1995–97). His interests include combinatorics, number theory, and representation theory.

GEORGE E. ANDREWS received his B.S. from Oregon State University in 1960, after which he spent a year at Cambridge University as a Fulbright Scholar. He was the late Professor Hans Rademacher's last student at the University of Pennsylvania, where he received his Ph.D. degree in 1964. His academic home base is the Pennsylvania State University, where he became an Evan Pugh Professor in 1981. In 1997 he was elected a Fellow of the American Academy of Arts and Sciences.

R. BRUCE RICHTER is interested in graph theory, combinatorics, elementary linear algebra in general contexts, and the topology of surfaces. He received his Ph.D. in 1983 from the University of Waterloo, Canada and has taught at Utah State University, The Ohio State University, and The U.S. Naval Academy. He has been at Carleton since 1988.

BILL WARDLAW received a B.A. in physics (Rice 1958), was Damage Control Assistant and Engineer Officer on USS HAMILTON COUNTY (LST 802) for two years, worked two years for Douglas Aircraft, and then got a Ph.D. in mathematics (UCLA 1966). He taught at the University of Georgia until going to the U.S. Naval Academy in 1972. Bill is interested in algebra, linear algebra, and elementary number theory. He lives with his wife and the youngest of his four children near Annapolis, Maryland.

PETER BRAZA received a Ph.D. in Applied Mathematics from Northwestern University in 1988. He is an associate professor at the University of North Florida. His research interests include nonlinear dynamics and chaos, particularly in optical devices such as the laser with injected signal. His classes have a relatively heavy physics flavor, reflecting his predilection towards modeling and applications. However, he enjoys forays out of his research area and into elementary "pure" problems like the one in the current paper.

JINGCHENG TONG was a high school teacher in China for ten years after graduating from Guizhou University. In 1985 he received a Ph.D. from Wayne State University and then went back to China and worked for two years at the Academia Sinica. He is a full professor at the University of North Florida. He has diverse research interests in many branches of mathematics and has published many papers.

TAKASHI AGOH was born in Shimane (Japan) and received both his B.S. degree and his Ph.D. in Mathematics from SUT (Science University of Tokyo). He is chairman of the Department of Mathematics at the Noda Campus of SUT. His research interests include mainly number theory and combinatorial theory. He enjoys jogging with pets, ballroom dancing, home carpentry, Sake testing, and Karaoke (singing to a taped accompaniment).

To call **PAUL ERDŐS** a phenomenon would be to understate his impact on mathematics as a research discipline and as a shared culture. Probably the greatest living icon of our subject, he still travels the globe, at the age of eighty-three, dropping in at universities and encouraging local researchers to new heights, spreading the word of the latest achievements and problems and, of course, writing joint papers. And then there's 'Erdős numbers', the number of co-authors between any given mathematician and Erdős. Einstein has an Erdős number of 2, Gauss has his Erdős number, as do thousands of mathematicians across the world. Twenty six different people have more than ten joint papers with Erdős, and two people have more than fifty. This is the first such collaboration for Agoh, and the second for Granville. *This biography was written by Granville, shortly before Professor Erdős passed away.*

ANDREW GRANVILLE was educated at Cambridge and Queens (Canada), with temporary positions at Toronto and the Institute for Advanced Study in Princeton, before coming to Georgia in 1991. He enjoys co-authoring papers, having had more than fifty co-authors, nine of whom he has never met thanks to the wonders of our electronic world. On the co-authoring hypergraph, there are two paths between him and Erdős of length one, at least 212 paths between them of length two, and several thousand of length three. He is an enthusiast for web-based information, and is slowly creating a website on *Arithmetic Properties of Binomial Coefficients*, which you may visit at
`http://www.math.uga.edu/~andrew/Binomial/index.html`

HUNG-HSI WU (Ph.D., MIT, 1963) is a differential geometer who knew nothing about school mathematics education prior to 1992. In that year, he stumbled upon the phenomenon known as the *mathematics education reform* when he innocently agreed to undertake the review of a new school mathematics curriculum (IMP). He was so taken aback that he proceeded to spend most of the next five years doing school mathematics education. Wu's publications include *General Relativity for Mathematicians* (with R. K. Sachs, 1977), and *The Bochner Technique in Differential Geometry* (1988).

JEREMY KILPATRICK received an A.B. and M.A. from UC Berkeley and an M.S. and Ph.D. (in mathematics education under E. G. Begle) from Stanford. He taught at Teachers College, Columbia, moving in 1975 to the University of Georgia, where he is Regents Professor of Mathematics Education. He is also Vice President of the International Commission on Mathematical Instruction. His interests include mathematics curricula, research in mathematics education, and the history of both.

JOHN L. DROST received his Ph.D. from the University of Miami in 1983 under the direction of E. H. Connell. He has been teaching at Marshall University since 1986 after a three-year stint at the University of Kentucky. His research interests are in algebraic number theory, combinatorics, and dynamical systems. His other interests are his wife Linda Hamilton, their son Jay, and Thumper, the big black bunny.

SCOTT AHLGREN is an Assistant Professor of Mathematics at Denison University. He enjoys doing research in number theory and riding a tandem bicycle with his wife Alison.

LARS ENGLISH is a graduate student in physics at Cornell University. The work in his Note was done while he was an undergraduate physics major at Denison University.

RONALD WINTERS holds the Tight Family Chair in the Physical Sciences at Denison University. The work in his Note stems from his interest in the use of computer-based symbolic manipulators in teaching mathematics and science.

SAUNDERS MAC LANE white, male, not yet dead, is said to have little practice in postmodern language, despite Sokal's hoax. He was President both of the MAA (1950–51) and of the AMS (1971–73); several other American mathematicians were equally fortunate. During WWII he directed research on the accuracy of airborn gyroscopic sights (for bombers over Germany); it is rumored that he and Eilenberg sneaked off evenings to do topology and discover categories. MacLane has carried out research on proof theory, algebraic numbers, planar graphs, group extensions, inseparable extensions of fields, cohomology of groups, hamiltonian mechanics, philosophy of mathematics, topos theory, you name it. At age 84 he apparently quit research in order to apply mathematical rigor to the current political attacks on mathematics and scientific research (as in his book manuscript, 500 pages). Stay tuned!

EZRA (BUD) BROWN is Professor of Mathematics at Virginia Tech. A mathematical grandson of L. E. Dickson via Gordon Pall, he has degrees from Rice and Louisiana State, and has been at Virginia Tech since 1969 with time out for sabbatical visits to Washington, DC and Munich. His main research interests are in number theory (especially computational) and graph theory, but he once wrote a paper with a sociologist. He enjoys teaching, singing in operas, and playing jazz piano. His wife Jo is teaching him to be a gardener. He is learning to cook—recipes for biscuits and for black beans are available by e-mail—and to juggle. The night sky has always held a fascination for him.

D. G. ROGERS graduated from Oxford in 1973, and has since traveled widely giving seminars; his personal record is six dozen seminars in one dozen countries in 1994. He was first in touch with the Howard University Combinatorics Group in 1976 and has visited several times, most recently in August and October 1995, when the authors broached the topic presented in this paper.

LOU SHAPIRO has been at Howard University since 1967. His mathematical interests involve generating functions, random walks, groups, and combinatorics, with the Catalan numbers being a frequent visitor in his 30 research articles. He is perhaps best known as a square dance caller, where his knowledge of dihedral group actions on a set of eight elements can safely be said to have been of no help whatsoever. His other interests include his family, music, detective novels, teaching, and coaching Howard's math team.

WEN-JIN WOAN received his Ph.D. from the University of Illinois under the direction of M. Suzuki, and has taught at Howard University since 1970. His main research areas are combinatorics and group theory.

THE MATHEMATICAL ASSOCIATION OF AMERICA

Numerology

or, What Pythagoras Wrought

Series: MAA Spectrum

Underwood Dudley

Underwood Dudley has done it again with a witty, fascinating book about number mystics. If you enjoyed Underwood Dudley's* Mathematical Cranks, *you must buy this book.

Underwood Dudley has assembled another delightful collection of essays that will amuse, engage and instruct you. Dudley, author of the immensely popular MAA titles *Mathematical Cranks,* and *The Trisectors,* has turned his attention in this volume to numerologists. Once you start reading about them, you won't be able to put the book down.

We learn in the introduction:

> "For some people, numbers do much more than merely count and measure. For some people, numbers have meanings, they have inwardness, they can be magic and versatile, or young and sprightly. I am not one of those people, since I think numbers have quite enough to do as it is, but for the crowd of number mystics, numerologists, pyramidologist, number-of-the-beasters, and others whose ideas and work will be described in the following chapters numbers have powers far out of the ordinary."

Number mystics, Dudley explains, originated with Pythagoras 2500 years ago and continue to this day. Numerology is applied number mysticism and is a more recent invention. You will find a history of number mysticism and numerology in the book, with a wealth of examples from the past as well as the present. Meet the Elliott Wave Theorists (who explain the movement of the stock market with Fibonacci numbers); the Bible-numberists who find 7s, 11s, 13s, or perfect squares in the Bible; the researcher who finds 57s throughout the American Revolution; the pyramidologists who see all of human history in numbers derived from measurements of the great pyramid of Egypt, and much more. Meet them all in the pages of this wonderful new book.

Catalog Code: NUMR/JR

332 pp., Paperbound, 1997, ISBN 0-88385-524-0

List: $29.95 MAA Member: $22.95

Phone in Your Order Now! ☎ *1-800-331-1622*

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders:** $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ______________________________

Address ______________________________

City ______________ State _____ Zip _______

Phone ______________________________

QTY.	CATALOG CODE	PRICE	AMOUNT
_______	NUMR/JR	_______	_______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

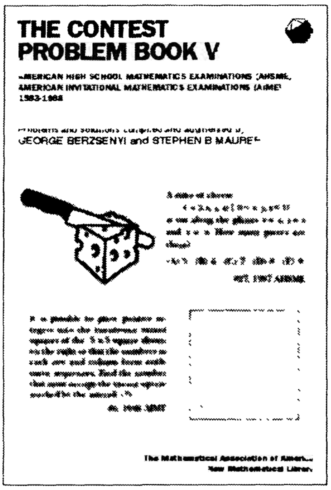
Shipping & handling _______

TOTAL _______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________________ Expires ___/___

Signature ______________________________



The Contest Problem Book V

American High School Mathematics Examinations and American Invitational Mathematics Examinations, 1983–1988

Series: New Mathematical Library

George Berzsenyi and Stephen B Maurer

Over the years perhaps the most popular of the MAA problem books have been the high school contest books, covering the yearly American High School Mathematics Examinations (AHSME) that began in 1950, co-sponsored from the start by the MAA. Book V also includes the first six years of the American Invitational Mathematics Examination (AIME) which was developed as an intermediate step between the AHSME and the USA Mathematical Olympiad (USAMO). The AIME has a unique answer format — all answers are integers between 0 and 999.

The editors of this volume, George Berzsenyi and Stephen B Maurer, were respectively the chair of the AIME and the AHSME during this period. In addition to a thorough index, they have added much material not included in Contest Books I–IV:

- a comprehensive guide to other problem materials world wide,
- additional solutions,
- dropped problems,
- statistical information,
- information on test development and history.

This volume is a must for avid fans of elementary problems.

Contest Books I–IV appear as NML volumes 5, 17, 25, and 29.

Catalog Code: NML-38/JR
308 pp., Paperbound, 1997
ISBN 0-88385-640-9
List: $24.95 MAA Member: $20.00

Phone in Your Order Now! ☎ 1-800-331-1622

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders**. $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ______________________

Address ______________________

City ____________ State _____ Zip ______

Phone ______________________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	NML-38/JR	______	______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling ______

TOTAL ______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________